Notas de aula de Econometria Aplicada
MEST 141

Prof. Daniel Santos

Setembro - Novembro 2008

1 Revisão

1.1 Causalidade

A teoria econômica investiga o comportamento decisório dos agentes econômicos. Em geral, os modelos produzem relações entre variáveis determinadas endógenamente e outras determinadas exogenamente

$$y_i = g_i\left(x_i, x_i^*, u_i\right)$$

O conteúdo empírico do modelo acima depende fundamentalmente da disponibilidade de dados. Na notação, y designa a variável endógena ao passo que x, x* e u, as exógenas. Por enquanto, vamos supor que y, x e x* são observadas e u, não. O efeito causal de uma mudança de x para x' em y é definido por

$$\Delta y_i = g_i\left(x_i', x_i^*, u_i\right) - g_i\left(x_i, x_i^*, u_i\right)$$

, ou seja, o efeito de uma alteração em x sobre y, mantendo-se constante as demais variáveis. Note que em princípio poderíamos ter um efeito diferente para cada indivíduo.

2 dificuldades surgem na descrição acima, e no fundo originam a maioria dos problemas econométricos descritos nos livros-texto. A primeira limitação é que na maioria dos casos, um mesmo indivíduo não é observado em dois estados distintos, x e x'. A segunda é que

mesmo que observássemos um mesmo indivíduo nesses dois estados, nada garante que a variável não observada u se manteve constante entre estas medições.

Para contornar estes problemas, costuma-se supor que (H1) as funções $g_i(.)$ são do tipo:

$$g_i(x_i, x_i^*, u_i) = \mu(x_i, x_i^*) + u_i$$

, e que (H2) a média de u não muda se repartirmos a amostra em diferentes grupos segundo os valores de x ($E(u|x) =$ constante).

$$\begin{aligned} E(\Delta y) &= E(y|x', x^*) - E(y|x, x^*) \\ &= E[\mu(x', x^*)|x, x^*] - E[\mu(x, x^*)|x, x^*] + E(u|x', x^*) - E(u|x, x^*) \\ &= \mu(x', x^*) - \mu(x, x^*) = g_i(x_i', x_i^*, u_i) - g_i(x_i, x_i^*, u_i) \end{aligned}$$

Há duas formas de tentar garantir H2: o método experimental (modelo de Rubin) e o método "econométrico" ou estrutural. Na primeira, o pesquisador varia aleatoriamente o nível de x para x' em um subgrupo da população e mede a variação correspondente em y. Por que isso funciona? Primeiramente, se originalmente tínhamos uma amostra homogênea em (x,x*) e a repartimos aleatoriamente em duas, a distribuição conjunta de (y,x,x*,u) em ambas deveria ser aproximadamente idêntica, de modo que se medíssemos as média de y e u em ambas, o resultado deveria ser semelhante. Como por hipótese as variáveis x,x* e u são determinadas fora do sistema, uma mudança controlada de x não deveria alterar a média de u. Os problemas nesse caso são que (i) raramente conseguimos realizar tal procedimento quando se trata de agentes e variáveis econômicas, (ii) mesmo que seja possível, dificilmente as condições laboratoriais necessárias para essa randomização espelharão as condições presentes

quando os agentes de fato tomam suas decisões, (iii) é possível que u varie com a mudança de x, se for também endógeno ao modelo, ainda que a mudança em x tenha sido exógena.

Quando a abordagem experimental não está disponível ou se revela inadequada, a alternativa é modelar explicitamente a relação entre x e u, e o sucesso da estimação dependerá crucialmente do realismo das hipóteses envolvidas. A hipótese mais comum é simplesmente E(u|x) = 0 (MQ-O). Alternativamente, pode-se argumentar que exista uma terceira variável z tal que mudanças em z alterem os valores de x sem alterar os de u (IV). O impacto nesse caso é semelhante ao experimental (o experimento pode ser visto como um caso particular de variável instrumental). Uma terceira alternativa é aceitar que H2 não vale, e modelar explicitamente o comportamento de $E\left(u|x', x^*\right) - E\left(u|x, x^*\right)$. A essa solução dá-se o nome de funções de controle, e em geral envolve uma interpretação rígida do significado de u e seu papel na determinação do comportamento dos agentes.

Nosso exemplo-canônico é a determinação do efeito causal de um ano a mais de escolaridade sobre salários. Suponha que nosso modelo teórico proponha que as pessoas vivam T anos, e passem S anos estudando e T-S trabalhando. Diferentes ocupações requerem diferentes níveis de escolaridade e remuneram de forma diferente os trabalhadores[1]. O valor presente do fluxo de renda de um indivíduo é, neste caso,

$$VP\left(S\right) = w(S)\int_S^T e^{-rt}dt = \frac{w\left(S\right)}{r}\left[e^{-rS} - e^{-rT}\right]$$

Em equilíbrio, os diferentes níveis educacionais deveriam proporcionar o mesmo valor presente de renda, o que significa que aqueles que estudaram mais precisam ser compensados

[1] São hipóteses do modelo que não há incertezas, que os trabalhadores são inicialmente idênticos em habilidades e oportunidades, e que os mercados de crédito são perfeitos.

pelo tempo em que ficaram sem receber salários:

$$\begin{aligned}
\frac{w(S)}{r}\left[e^{-rS}-e^{-rT}\right] &= \frac{w(0)}{r}\left[1-e^{-rT}\right] \\
\ln w(S) &= \ln w(0)+rS+\ln\left[\frac{1-e^{-rT}}{1-e^{r(T-S)}}\right] \\
&\approx \ln w(0)+rS, \text{ se T for grande}
\end{aligned}$$

Nesse modelo, os indivíduos são em princípio indiferentes entre os diversos níveis de escolaridade, e serão distribuídos aleatoriamente entre os vários níveis de S, sugerindo uma relação do tipo:

$$\begin{aligned}
\ln w_i &= \ln w(0)+rS_i+\varepsilon_i \\
&\quad \varepsilon_i \text{ independente de } S_i
\end{aligned}$$

Como o erro da equação acima é puramente aleatório, podemos consistentemente estimar uma regressão do tipo:

$$\ln w_i = \beta_0+\beta_1 S_i+\varepsilon_i$$

por MQO numa cross-section, e nesse caso a interpretação do coeficiente β_0 é a de que representaria o (log do) salário de um analfabeto, e β_1 o retorno r obtido pelo investimento em educação, que nesse caso também representa o efeito causal de S sobre $\ln(w)$.

O modelo alternativo é um mercado de trabalho onde as firmas usam características produtivas dos agentes como insumos para produzir bens, segundo a função de produção

$$Q = Q(S, x^*, u) = \exp\left[b_0+b_1 S+b_2 u\right]$$

Em equilíbrio, o salário será $w = w^s + w^u = (b_1 + b_2)\, Q$, o que implica:

$$\begin{aligned} \ln w &= [b_0 + \ln(b_1 + b_2)] + b_1 S + b_2 u \\ &= \beta_0 + \beta_1 S_i + u_i \end{aligned}$$

Note que enquanto no primeiro modelo o termo ε_i era puramente aleatório (implicando em que $\mathrm{E}(\varepsilon|x) = 0$ seja razoável), no segundo, o termo u_i reflete uma característica não observável do indivíduo e que pode estar correlacionada com escolaridade (e.g. habilidade).

Um terceiro modelo (Mincer, 1974) propõe ainda uma terceira forma relacionando salários e educação:

Seja E_t = salário potencial no período t, w_t = salário observado, C_t = investimento em capital humano, k_t = fração do salário potencial investida, ρ_t = retorno ao investimento realizado em capital humano.

$$\begin{aligned} C_t &= kE_t \\ E_{t+1} &= E_t + \rho_t C_t \\ &= (1 + k_t \rho_t)\, E_t \\ &= \prod_{j=0}^{t} \left(1 + k_j \rho_j\right) E_0 \end{aligned}$$

Suponha que na educação formal o investimento em capital humano é máximo: $k_s = 1$,

e defina ρ_s = retorno à educação e ρ_0 = retorno ao treinamento pós-educação. Então[2][3]:

$$\begin{aligned}\ln E_t &= \ln E_0 + s\ln(1+\rho_s) + \rho_0 \sum_{j=s}^{t-1} \ln(1+\rho_0 k_j) \\ &\approx \ln E_0 + s\rho_s + \rho_0 \sum_{j=s}^{t-1} k_j\end{aligned}$$

Finalmente, e supondo que o investimento em aprendizado decresce a uma taxa linear após concluídos os estudos:

$$\begin{aligned}k_{s+x} &= \kappa\left(1-\frac{x}{T}\right) \\ \ln E_{s+x} &\approx \ln E_0 + \rho_s S + \left(\kappa\rho_0 + \frac{\kappa\rho_0}{2T}\right)x - \frac{\kappa\rho_0}{2T}x^2\end{aligned}$$

[2] Expandindo $\ln(x)$ em Taylor em torno de 1:

$$\ln(x) = \ln(1) + (x-1)\frac{1}{1} - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} + ...$$

a expansão em 1^a ordem fornece a aproximação sugerida (note que $\ln(1) = 0$).

[3]

$$\begin{aligned}z &= j - s \\ \sum_{j=s}^{x-1} k_j &= \sum_{z=0}^{x-1} k_{s+z} \\ &= \kappa\sum_{z=0}^{x-1} 1 - \frac{\kappa}{T}\sum_{z=1}^{x-1} z \\ &= \kappa x - \frac{\kappa}{T}\frac{(x-1)}{2}\left[2 + (x-2)\,1\right] \\ &= \kappa x - \frac{\kappa}{T}\frac{(x-1)\,x}{2}\end{aligned}$$

e como

$$\begin{aligned} w_{s+x} &= (1-k_{s+x})\,E_{s+x} \\ \ln w_{s+x} &\approx \ln E_{s+x} - \kappa\left(1-\frac{x}{T}\right) \\ &= (\ln E_0 - \kappa) + \rho S + \left(\kappa\rho_0 + \frac{\kappa\rho_0}{2T} + \frac{\kappa}{T}\right)x - \frac{\kappa\rho_0}{2T}x^2 \\ &= \beta_0 + \beta_1 S + \beta_2 x + \beta_3 x^2 \end{aligned}$$

Note que agora para estimar o efeito de educação sobre salários, é preciso incluir também experiência em nossa regressão.

1.2 Estimação

1.2.1 Minimização de função perda (ex. MQO)

Seja o modelo $y_i = \mu(x_i, x_i^*) + u_i$, suponha que $E(u|x) = 0$, e defina como perda a função $L[u]$. Por exemplo, se nossa intenção for a de explicar ao máximo a variação de y com a variação de x,x*, então podemos decidir que $L(u) = u^2$, pois esta função atinge o mínimo precisamente quando u = 0. Nessa estratégia, o modelo ideal deverá resolver $\min_\mu E(u^2) = \min_\mu E\left[(y-\mu(x,x^*))^2\right]$, e o modelo estimado, $\min_\mu \sum_{i=1}^N (y_i - \mu(x_i, x_i^*))^2$. Note que um caso particular de $\mu(x_i, x_i^*)$ é $\mu(x_i, x_i^*) = b_0 + b_1 x_i + b_2 x_i^*$. Se supor essa forma para a função $\mu(x_i, x_i^*)$ for justificável, então a solução será MQO.

1.2.2 Maximização da função de verossimilhança

Considere o mesmo modelo acima, mas agora suponha que estejamos dispostos a assumir que o termo não observado, u, seja distribuído segundo uma função de densidade de probabilidade $f(.)$ e independente de x. Suponha que você escolheu um valor $\widehat{\beta}^*$ para ser seu

estimador. Numa amostra com $i = 1, ..., N$ indivíduos, a probabilidade de que o resíduo de um determinado indivíduo venha da distribuição f é:

$$\begin{aligned} \widehat{u}_i^* &= y_i - x_i\widehat{\beta}^* \\ \Pr\left(u_i = \widehat{u}_i^*\right) &= f\left(\widehat{u}_i^*\right) \end{aligned}$$

Numa amostra aleatória (onde as realizações de u_i para diferentes indivíduos são independentes), temos que a probabilidade de que todos os resíduos da amostra venham da distribuição $f(.)$ é:

$$\begin{aligned} \Pr\left(u_1 = \widehat{u}_1^*, u_2 = \widehat{u}_2^*, ..., u_N = \widehat{u}_N^*\right) &= \Pr\left(u_1 = \widehat{u}_1^*\right) x \Pr\left(u_2 = \widehat{u}_2^*\right) x ... x \Pr\left(u_N = \widehat{u}_N^*\right) \\ &= \prod_{i=1}^{N} f\left(u_i\right) = \prod_{i=1}^{N} f\left(y_i - x_i\beta\right) \end{aligned}$$

A estratégia aqui é escolher o valor de β que maximiza a função acima, pois esse é o valor que maximiza a probabilidade de que seus resíduos estimados, $\widehat{u}_i^*$, venham de fato da distribuição f. É fácil mostrar que se uma função $g(\beta)$ atinge seu máximo em β^*, então a função $\ln g(\beta)$ também atinge seu máximo em β^*. Como é mais fácil para os computadores maximizar uma soma do que um produto, convenciona-se usar o logarítmo da verossimilhança ao invés da própria verossimilhança na estimação.

Um exercício interessante é mostrar que se nossa hipótese é a de que $f(u)$ é a densidade da distribuição normal (ou gaussiana), então $\widehat{\beta}^{MV} = \widehat{\beta}^{MQO}$.

1.2.3 Método dos momentos

Suponha novamente que $y_i = x_i\beta + u_i$, e acrescente as hipóteses de que $E(u) = 0$ e $E(xu) = 0$. Como o primeiro elemento de x é a constante 1, a primeira condição acima é simplesmente

$E(1 * u) = 0$, e já está portanto incorporada em $E(xu) = 0$. Sabemos que um estimador consistente dos momentos populacionais acima é o equivalente amostral:

$$\sum_{i=1}^{N} x_i (y_i - x_i\beta)' = 0$$

que implica em:

$$\sum_{i=1}^{N} x_i y_i - x_i x_i'\beta = 0$$

$$\beta = (x'x)^{-1} x'y = \widehat{\beta}^{MQO}$$

Genericamente, o econometrista supõe uma série de hipóteses sobre os momentos populacionais dos resíduos, e usa os equivalentes amostrais para encontrar o estimador.

Suponha no problema acima que ao invés de $E(xu) = 0$ fosse suposto que $E(u|x) = 0$. Ora, primeiramente temos que $E(u|x) = 0$ implica em $E(xu) = 0$, mas também implica que para qualquer função $g(x)$, $E(g(x)u) = 0$, o que leva em princípio a uma infinidade de condições sobre momentos amostrais (por exemplo, $E(x^2u) = 0$, $E(\cos(x)u) = 0$, etc). A pergunta é então, como lidar com esse excesso de condições? A resposta leva ao Método Generalizado dos Momentos, que consiste nos seguintes passos:

1. construir os equivalentes amostrais para cada momento que o analista considere relevante. No exemplo do parenteses acima (e considerando apenas uma variável explicativa

x):

$$\begin{aligned} E(u) &= 0 : \sum_{i=1}^{N} (y_i - x_i\beta) = 0 \\ E(xy) &= 0 : \sum_{i=1}^{N} x_i (y_i - x_i\beta) = 0 \\ E\left(x^2 y\right) &= 0 : \sum_{i=1}^{N} x_i^2 (y_i - x_i\beta) = 0 \\ E(\cos(x) u) &= 0 : \sum_{i=1}^{N} \cos(x_i) (y_i - x_i\beta) = 0 \end{aligned}$$

2. escolher uma matriz positiva (semi) definida M, para dar pesos aos J diferentes momentos. Se a matriz for diagonal, o peso de cada condição j na estimação final será dado por $m_j/(m_1 + m_2 + ... + m_J)$.

3. Empilhar os momentos num vetor

$$V = \begin{bmatrix} \sum_{i=1}^{N} (y_i - x_i\beta) \\ \sum_{i=1}^{N} x_i (y_i - x_i\beta) \\ \sum_{i=1}^{N} x_i^2 (y_i - x_i\beta) \\ \sum_{i=1}^{N} \cos(x_i) (y_i - x_i\beta) \end{bmatrix}$$

4. Escolher β que minimize o critério $V'MV$

Nessa estratégia, qualquer matriz positiva definida M que o econometrista escolha produzirá uma estimativa consistente de β. Há contudo critérios para escolher de forma ótima a matriz M, de modo a minimizar a variância do estimador. Mais importante do que as tecnicalidades é convencer e ser convencido de que a hipótese $E(u|x) = 0$ (ou qualquer outra que seja necessária à estimação) é realística do ponto de vista econômico.

1.3 Identificação

Dizemos que um modelo é identificado dentro de uma classe de modelos se, dadas as implicações empíricas derivadas das hipóteses do modelo, nenhum outro membro desta classe é observacionalmente equivalente ao modelo referido. Em nosso exemplo, suponha que eu observo o logarítmo dos salários e a educação, e que as hipóteses do modelo sejam:

$$\begin{aligned} \ln w_i &= \beta_0 + \beta_1 S_i + \varepsilon_i \\ E(\varepsilon) &= 0 \\ E(S\varepsilon) &= 0 \end{aligned}$$

Dadas essas hipóteses, temos que

$$\begin{aligned} E(\ln w) - \beta_0 - \beta_1 E(S) &= 0 \\ E(\ln w) - \beta_1 E(S) &= \beta_0 \end{aligned}$$

$$\begin{aligned} E[(\ln w - \beta_0 - \beta_1 S) S] &= 0 \\ E\left[\left(S \ln w - \beta_1 S^2 - (E(\ln w) - \beta_1 E(S)) S\right)\right] &= 0 \\ E[S(\ln w - E(\ln w))] - \beta_1 E[S(S - E(S))] &= 0 \\ \frac{E[S(\ln w - E(\ln w))]}{E[S(S - E(S))]} &= \beta_1 \end{aligned}$$

e finalmente:

$$\begin{aligned} \beta_1 &= \frac{cov(S, \ln w)}{var(S)} \\ \beta_0 &= E(\ln w) - \frac{cov(S, \ln w)}{var(S)} E(S) \end{aligned}$$

Note que todos os parâmetros do modelo foram expressos como funções de momentos amostrais, que não dependem de parâmetro algum. Isso significa que se eu tivesse toda a população em mãos, poderia simplesmente medir (unicamente) os valores de β_0, β_1, e nenhuma outra combinação destes parâmetros satisfaria as hipóteses do modelo. Dizemos portanto que β_0, β_1 são identificados na classe de modelos que satisfaz tais hipóteses.

1.4 Teoria assintótica

- Lei forte dos grandes números (LFGN): Seja $X_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$, e $X = \lim_{N\to\infty} E(X_N)$. Se X_N converge "quase certamente" para X (isto é, $X_N \xrightarrow{a.s.} X$), então dizemos que $\{Y_i\}$ obedece LFGN. 2 conjuntos de condições suficientes para isso são:

$$\{Y_i\} \text{ independentemente distribuídos e } \sum_{i=1}^{N} \frac{1}{i^2} var(Y_i) < \infty$$

ou

$$\{Y_i\} \text{ i.i.d. e } E(Y_i) \text{ existe}$$

- Lei forte dos grandes números (LFGN): Seja $X_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$, e $X = \lim_{N\to\infty} E(X_N)$. Se X_N converge em probabilidade para X, (isto é, $X_N \xrightarrow{p} X$), então dizemos que $\{Y_i\}$ obedece LFGN. Condição suficiente para isso é:

$$\lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} var(X_N) = 0$$

- Teorema de Slutsky: se $X_N \xrightarrow{p} X$ e $g(X)$ é uma função contínua, então $g(X_N) \xrightarrow{p} g(X)$.

- Convergência em distribuição: $X_N \xrightarrow{d} X$ se e somente se $F(X_N) \xrightarrow{d} F(X)$ ponto a ponto, onde $F(.)$ é a distribuição acumulada de X.

- Teoremas do Limite Central:

  (i) Lindberg-Levy: $\{Y_i\}$ i.i.d.; $\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$; $E(Y_i) = \mu$; $var(Y_i) = E(Y_i - \mu)^2 = \sigma^2$. Então: $\sqrt{N}\left(\overline{Y} - \mu\right) \xrightarrow{d} N(\mu, \sigma^2)$

  (ii) Liapounov: $\{Y_i\}$ independentemente distribuído, $E(Y_i) = \mu$; $var(Y_i) = \sigma_i^2$; $E(Y_i^3)$ existe. Então $\sqrt{N}\left(\overline{Y} - \mu\right) \xrightarrow{d} N(\mu, \overline{\sigma}^2)$, onde

$$\overline{\sigma}^2 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \sigma_i^2$$

- Corolários de Slutsky com TLC:

  (i) $g(.)$ contínua e $X_N \xrightarrow{d} X$ então: $g(X_N) \xrightarrow{d} g(X)$

  (ii) $X_N \xrightarrow{p} Y_N$ e $Y_N \xrightarrow{d} Y$, então $X_N \xrightarrow{d} Y$

- Teoremas de Wald e Man: seja $X_N \xrightarrow{d} X$ e $Y_N \xrightarrow{p} c$ (constante). Então:

  (i) $X_N + Y_N \xrightarrow{d} X + c$

  (ii) $X_N Y_N \xrightarrow{d} Xc$

  (iii) O limite da distribuição conjunta de (X_N, Y_N) existe e é igual à distribuição conjunta de (X_N, c)

- Comparando conceitos de convergência: $X_N \xrightarrow{a.s.} X \Rightarrow X_N \xrightarrow{p} X \Rightarrow X_N \xrightarrow{d} X$

- Método Delta (escalar): seja α_N uma sequência determinística de números tendendo a infinito (ex: $\sqrt{N}$). Suponha que $\alpha_N (Y_N - c) \xrightarrow{d} X$, onde c é uma constante. Considere uma função contínua e duas vezes diferenciável, $g(.)$. Então:

$$g(Y_N) \approx g(c) + g'(c)(Y_N - c)$$

$$\begin{aligned} \alpha_N [g(Y_N) - g(c)] &\approx g'(c)[\alpha_N (Y_N - c)] \\ &\xrightarrow{d} g'(c) X \end{aligned}$$

em particular, se $\sqrt{N}(Y_N - c) \xrightarrow{d} X \sim N(0, \sigma^2)$, então $\sqrt{N}[g(Y_N) - g(c)] \xrightarrow{d} g'(c) X \sim N\left(0, [g'(c)\sigma]^2\right)$

- Se $g(.)$ e X são multidimensionais, e $X \sim N(0, \Sigma)$, então o método Delta se estende para

$$\sqrt{N}[g(Y_N) - g(c)] \xrightarrow{d} N\left[0, E\left(\nabla_g(Y) \Sigma \nabla_g'(Y)\right)\right]$$

- Aplicação em máxima verossimilhança: suponha que $\ln \pounds(\beta)$ é tal que no ótimo, o parâmetro β_0 maximiza $\ln \pounds(\beta)$, isto é:

$$\begin{aligned} E\left(\frac{\partial \ln \pounds(y_i; \beta_0)}{\partial \beta}\right) &= E(s_i(y; \beta_0)) = 0 \\ E\left(\frac{\partial^2 \ln \pounds(\beta_0)}{\partial \beta \partial \beta'}\right) &= E(H_{i\pounds}(y_i; \beta_0)) : \text{negativo-definido} \end{aligned}$$

Defina os correspondentes amostrais:

$$\begin{aligned} s(y; \beta) &= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ln \pounds(y_i; \beta)}{\partial \beta} \\ H_{\pounds}(y; \beta) &= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \ln \pounds(y_i; \beta)}{\partial \beta \partial \beta'} \end{aligned}$$

então, se $s\left(y;\widehat{\beta}\right) \xrightarrow{p} s\left(y;\beta_0\right) \xrightarrow{p} E\left(s\left(y;\beta_0\right)\right)$ e $H_{\pounds}\left(y;\widehat{\beta}\right) \xrightarrow{p} H_{\pounds}\left(y;\beta_0\right) \xrightarrow{p} E\left(H_{i\pounds}\left(y_i;\beta_0\right)\right)$, temos

$$\begin{aligned}
\sqrt{N}s\left(y;\beta_0\right) &= \sqrt{N}\frac{1}{N}\sum_{i=1}^{N} s_i\left(y_i;\beta_0\right) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} s_i\left(y_i;\beta_0\right) \xrightarrow{d} N\left[0,A_0\right] \\
A_0 &= E\left(s\left(y;\beta_0\right)s\left(y;\beta_0\right)'\right)
\end{aligned}$$

$$s\left(y_i;\widehat{\beta}\right) = \frac{1}{N}\sum_{i=1}^{N} s_i\left(y_i;\widehat{\beta}\right) = 0 \text{ (FOC)}$$

$$\begin{aligned}
s\left(y_i;\widehat{\beta}\right) &\approx s\left(y_i;\beta_0\right) + H_{\pounds}\left(y;\beta_0\right)\left(\widehat{\beta}-\beta_0\right) \\
\sqrt{N}\left(\widehat{\beta}-\beta_0\right) &\approx \left[-H_{\pounds}^{-1}\left(y;\beta_0\right)\right]\left[\sqrt{N}s\left(y_i;\beta_0\right)\right] \xrightarrow{d} N\left[0,B_0A_0B_0\right] \\
B_0 &= -E\left[H_{\pounds}^{-1}\left(y;\beta_0\right)\right]
\end{aligned}$$

2 Escolha discreta - modelo de utilidades aleatórias (inspirado nas notas de aula de J.J. Heckman)

A economia tem como objeto de estudo o comportamento decisório dos agentes, que tipicamente escolhem a melhor ação a ser tomada dentre as diversas possibilidades que compõem o conjunto de escolha dos indivíduos. Dessa forma, grande parte das predições testáveis dos modelos econômicos é obtida através da análise das soluções dos problemas de maximização dos agentes, que tipicamente se resumem em condições de primeira e segunda ordem.

No caso de decisões de consumo, por exemplo, um indivíduo com utilidade definida pelo parâmetro de preferências U, com renda Y e que observa preços P, escolhe quando consumir

de um bem g resolvendo:

$$\max_{g,C} u\left(g, C; U\right)$$

$$s.t. \quad : \quad p^g g + p^c C \leq Y$$

e a solução deste problema $g^* = g^*\left(Y, p^g, p^c, U\right)$ fornece a demanda pelo bem g por parte do indivíduo. Neste caso, as variáveis (Y, p^g, p^c, U) são determinadas fora do problema de otimização e assumidas como dadas pelo agente, enquanto g é a variável determinada endogenamente. Supondo que a função $g^*\left(Y, p^g, p^c, U\right)$ possa ser rescrita como $g_0^*\left(Y, p^g, p^c\right) + U$, com U independente de Y, p^g, p^c, e que o econometrista tenha acesso a uma base de dados com informações sobre $\left(g^d, Y, p^g, p^c\right)$, é possível então estimar o modelo comportamental acima regredindo g^d em (Y, p^g, p^c). Note neste exemplo que (1) há clara justificativa teórica para a hipótese de que as variáveis (Y, p^g, p^c) podem ser classificadas como exógenas no exercício estatístico, pois (segundo o modelo) os agentes econômicos não tem meios de afetar seus valores, (2) há clara justificativa teórica para a fonte de aleatoriedade do modelo (parâmetro não observado de preferências que varia entre indivíduos), (3) a independência entre o termo não observável e as demais variáveis pode ser justificada (geralmente a etapa mais controversa de qualquer exercício), (4) as derivadas parciais $\partial g_0^* / \partial (Y, p^g, p^c)$ podem ser interpretadas, em se aceitando as hipóteses do modelo, como efeitos causais de mudanças nestas variáveis sobre a demanda por g.

Suponha agora que um indivíduo com características observáveis, X, e não-observáveis, U, escolhe de maneira ótima a ação $a^* \in A$. Caso a represente uma variável contínua (isto é, caso A seja denso), a solução pode ser representada genericamente como $a^* = a^*\left(X, U\right)$. No

entanto, dois problemas podem surgir, cuja solução é parecida. Primeiramente, o conjunto de escolha pode ser composto por um número finito e discreto de alternativas (por exemplo, decidir se aceito uma proposta de trabalho da firma A ou da firma B, se caso com F ou com R, se viajo para E ou para O, etc.). A segunda possibilidade é que a seja contínua, mas que observemos apenas um indicador discriminando se $a^* \gtrless \overline{a}$ (por exemplo, um sujeito recebe várias propostas de emprego e tem salário de reserva W. Ele deve aceitar a proposta que pagar mais, mas se em nossos dados não houver informação explícita sobre o salário contratado, saberemos apenas que $w \geq W$).

2.1 Modelo de utilidades aleatórias:

Suponha que $A = \{a_1, ..., a_J\}$. Cada alternativa a_j tem associada um bem-estar $u(a_j; X, U)$, que em princípio pode variar para pessoas com diferentes características X, U (tanto X quanto U representam vetores, isto é, $X = [1, X_1, ..., X_K]$ e $U = [U_1, ..., U_J]$ de atributos individuais que afetam as decisões dos agentes). Defina o conjunto de variáveis indicadoras $\{d_j\}_{j=1}^{J}$ tal que $d_j = 1$ se a alternativa j é escolhida e $d_j = 0$ caso contrário. Como nosso interesse reside no caso em que as alternativas são mutuamente excludentes, isto é, em que os agentes podem escolher apenas uma ação dentre as J permitidas, podemos caracterizar o problema de otimização como:

$$\max_{\{d_j\}_{j=1}^{J}} \sum_{j=1}^{J} d_j u(a_j; X, U)$$
$$s.t. \quad : \quad \sum_{j=1}^{J} d_j = 1$$

Dado o comportamento maximizador dos indivíduos, a alternativa eleita terá que satis-

fazer a condição:

$$d_j = 1 \Leftrightarrow u\left(a_j; X, U\right) > u\left(a_k; X, U\right); \forall k \neq j : a_k \in A$$

Em palavras, o indivíduo escolherá j se e somente se esta alternativa proporcionar um nível de bem-estar maior que qualquer outra das possibilidades.

Suponha que as variáveis não observáveis, U, são compostas por um vetor de J escalares, $U_1, ..., U_J$, e que apenas a coordenada j afeta o nível de utilidade correspondente a a_j, que além disso é linearmente separável em X e U, isto é,

$$u\left(a_j; X, U\right) = u\left(a_j; X, U_j\right) = v_j\left(X\right) + U_j$$

Aceitando estas hipóteses como verdadeiras, a regra decisória fica:

$$\begin{aligned} d_j &= 1 \Leftrightarrow v_j\left(X\right) + U_j > v_k\left(X\right) + U_k; \forall k \neq j : a_k \in A \\ &\Leftrightarrow U_k - U_j < v_j\left(X\right) - v_k\left(X\right) \end{aligned}$$

Nossos dados são compostos por N indivíduos para os quais observamos características X e decisões, $\{d_j\}_{j=1}^{J}$. No caso geral, se estivermos dispostos a assumir alguma forma funcional para a distribuição de $U_k - U_j$, então conseguimos estimar a função $H\left(X\right) = v_j\left(X\right) - v_k\left(X\right)$ (mas não podemos separar $v_j\left(X\right)$ de $v_k\left(X\right)$). Nesse ponto, três comentários são úteis: (1) dependendo da pergunta que se quer responder, a estimação de $H\left(X\right)$ é suficiente para responder. Em particular, se nosso interesse é em predizer como a proporção de pessoas que escolhe a alternativa j varia quando alguma característica X é modificada (e.g. como

a fração de consumidores de pepsi varia com a renda), não é necessário saber $v_j(X)$; (2) assim como na teoria do consumidor, transformações monotônicas das utilidades não afetam as decisões. Em particular, $c(U_k - U_j) < cH(X)$ para qualquer constante $c > 0$. Isso significa que a escala da variável não observável $\varepsilon = (U_k - U_j)$ tem que ser normalizada na estimação (isto está implícito na exigência de que o econometrista defina a distribuição de ε); (3) eventualmente observamos não apenas as decisões mas também as recompensas associadas às diferentes ações disponíveis. Com mais informação, pode-se em geral estimar mais parâmetros do que com menos informação, e no caso mais abrangente, podemos inclusive estimar separadamente $v_j(X), v_k(X)$ e elementos da distribuição de ε (exemplo: modelo de Roy).

2.1.1 Caso binário

A grande maioria dos exercícios empíricos lida com situações onde os indivíduos têm apenas duas ações possíveis em seus conjuntos de escolha, digamos, $A = \{0, 1\}$. Suponha que a parte explicada por características observáveis do benefício líquido associada a cada ação possa ser escrita como $v_j(X) = X\gamma_j$. O modelo então fica:

$$u(0; X, U) = X\gamma_0 + U_0$$

$$u(1; X, U) = X\gamma_1 + U_1$$

$$d = 1 \Leftrightarrow X(\gamma_1 - \gamma_0) > \varepsilon$$

Podemos também definir $\beta = (\gamma_1 - \gamma_0)$, e a probabilidade de que um indivíduo seja

observado adotando uma determinada ação é:

$$\begin{aligned} \Pr(d=1|X) &= \Pr[\varepsilon < X\beta] \\ \Pr(d=0|X) &= \Pr[\varepsilon \geq X\beta] \\ &= 1 - \Pr[\varepsilon < X\beta] \end{aligned}$$

Por definição, a função de distribuição acumulada de uma variável aleatória Z pode ser escrita como $F(z) = \Pr(Z < z)$, de modo que nosso modelo pode também ser representado como

$$\begin{aligned} \Pr(d=1|X) &= F(X\beta) \\ \Pr(d=0|X) &= 1 - F(X\beta) \end{aligned}$$

Qual a interpretação dos coeficientes $\beta = [\beta_0, \beta_1, ..., \beta_K]$ nesse caso? Considere uma característica específica, k. O termo $\beta_k X_k$ captura a parcela do ganho líquido de bem estar entre as alternativas 1 e 0 que pode ser atribuído à característica X_k.

Para ilustrar, suponha que um indivíduo tenha que decidir se passa as férias em Veneza (alternativa 0) ou na Disney (alternativa 1), e que X_k seja a idade. É de se esperar que pessoas mais velhas tenham ganho líquido negativo de bem estar de passar da alternativa 0 para a alternativa 1, controlado por outras características, e portanto β_k deveria ser negativo. Isso significa que TODAS as pessoas acima de uma certa idade optarão por ir a Veneza enquanto as que estão abaixo desse limite irão à Disney? A resposta é não. O ganho líquido de ir a Veneza com a idade pode ser mais do que compensado pela perda líquida de escolher essa ação devido a outras características do agente, incluindo-se aí as características não

observáveis, e por isso podemos acomodar o fato que provavelmente surgiria nos dados de que algumas pessoas mais velhas escolhem ir à Disney enquanto algumas crianças preferem Veneza.

Na próxima subseção falaremos um pouco de outra classe de modelos que têm a mesma estrutura empírica do visto acima. Em seguida estudaremos como estimar estes modelos e avaliar a qualidade de nossas estimações

2.2 Modelos de funções-índice

A motivação para esta classe de modelos é a seguinte: suponha que um indivíduo tenha um benefício constante garantido se não adotar uma ação (outside option, em inglês), e um benefício $B(X) + \varepsilon$, se adotar esta ação. Um exemplo é a proposta de emprego mencionada anteriormente, onde W pode representar o salário de reserva do trabalhador ou um seguro-desemprego que o agente só recebe se não estiver trabalhando, e w for o salário incluído na proposta. Digamos que o salário dependa de características observável, tais como educação e experiência, e um termo aleatório independente dos observáveis:

$$\ln w = X\beta + \varepsilon$$

Se o salário for observado, poderíamos regredir diretamente $\ln w$ em X e estimar desse modo os coeficientes β que nos interessam, mas o problema aqui é que nossa base de dados não tem salários, mas apenas a decisão dos indivíduos de aceitar ou não uma determinada proposta de trabalho, numa amostra de indivíduos que receberam proposta em um dado período. Um outro exemplo interessante é a estimação de uma curva de demanda por um

dado produto, onde sabemos apenas que, aos preços vigentes, os consumidores compram se e somente se os preços estiverem abaixo de sua propensão a pagar. Mas como a propensão a pagar por um bem não é observável, temos que nos satisfazer com a informação restrita ao fato de que os indivíduos compraram ou não o bem.

Num cenário bastante simples de maximização, podemos supor que os indivíduos aceitaram a oferta se e somente se

$$\ln w \geq \ln W$$

ou seja:

$$\begin{aligned} d &= 1 \Leftrightarrow X\beta + \varepsilon \geq \ln W \\ \Pr(d = 1|X) &= 1 - F(\ln W - X\beta) \end{aligned}$$

No modelo acima, o intercepto estimado representará $\ln W - \beta_0$, e todos os demais parâmetros $\beta_1, \beta_2, ..., \beta_K$ podem ser estimados. Qual a perda então em relação à situação em que os salários também constam da base de dados? Como foi dito, na situação descrita acima fomos forçados a assumir uma forma funcional específica para a distribuição do erro, ε, enquanto antes tínhamos apenas que supor que sua média era zero (além da independência entre ε e X). Isso porque, de novo, as decisões observadas são preservadas se transformarmos monotonicamente as utilidades de todas as alternativas envolvidas. No exemplo acima, o benefício de aceitar uma proposta é w, enquanto o benefício de não aceitar é W. O próprio logaritmo que incluímos é em si uma transformação que não muda a regra decisória. Considere uma multiplicação de todos os termos da desigualdade acima por 3. Como a desigualdade seria preservada, um modelo com coeficientes $\widetilde{\beta} = 3\beta$, $\widetilde{\varepsilon} = 3\varepsilon$ e $\widetilde{\ln W} = 3 \ln W$

racionalizaria os dados tão bem quanto nosso modelo original, e não seria possível discriminar qual o modelo que nos interessa. Ao supor que a distribuição de ε é fixa e dada por $F(\varepsilon)$, automaticamente acabamos com esse problema, pois a distribuição de $\widetilde{\varepsilon}$ não atenderia à hipótese de que a verdadeira distribuição do modelo é $F(.)$. Num contexto específico em que $F(\varepsilon)$ é suposta como sendo a distribuição normal com média 0, o modelo seria identificado se também fixássemos a variância (em geral assumindo $\sigma = 1$). Como a variância de uma normal é o parâmetro que mede sua escala, sendo sensível exatamente a operações multiplicativas da variável aleatória em questão, diz-se que em modelos com variável dependente binária os coeficientes β são identificados a menos da escala.

2.3 Estimação

2.3.1 Por que $\Pr(d = 1|X)$?

Tanto no modelo de utilidades aleatórias quanto no de funções-índice, a variável dependente é discreta, e no caso binário, tem a estrutura

$$\Pr(d = 1|X) = F(X\beta)$$

Por quê então nos preocupamos com $\Pr(d = 1|X)$? Simplesmente porque

$$\begin{aligned} E(d|X) &= E(d|X, d = 1)\Pr(d = 1|X) + E(d|X, d = 0)\Pr(d = 0|X) \\ &= 1 * \Pr(d = 1|X) + 0 * \Pr(d = 0|X) \\ &= \Pr(d = 1|X) \end{aligned}$$

Em outras palavras, a regressão (regressão = $E(y|X)$ por definição) de d em X fornece exatamente $F(X\beta)$. No caso geral, e diferentemente do que foi aprendido no caso de modelos

lineares, essa regressão não precisa necessariamente ser linear nas covariadas. Há inclusive motivos fortes para não desejar que seja! Consequentemente, os métodos de projeção utilizados para estimar modelos lineares (MQO, MQOG) raramente se aplicam nesses casos. Os métodos mais utilizados são os de máxima verossimilhança e mínimos quadrados não-lineares, e ambos requerem inicialmente a especificação da função $F(.)$ para serem estimados

2.3.2 Especificações comuns

Quatro especificações dominam a literatura:

(a) Probabilidades lineares (MPL): $\Pr(d=1|X) = X\beta$. Nessa especificação, a distribuição do termo aleatório ε é suposta como sendo uma uniforme (0,1), que tem média 1/2 (e não 0, como estamos acostumados). Alternativamente, diz-se apenas que d possui uma distribuição de Bernoulli com parâmetro $p = X\beta$.

(b) Probit: $\Pr(d=1|X) = \Phi(X\beta) = \int_{-\infty}^{X\beta} \frac{z}{\sqrt{2\pi}} \exp -\frac{z^2}{2} dz$. Nesse caso, ε tem distribuição Normal padrão. Interessante notar que no caso do modelo de utilidades aleatórias, tínhamos $\varepsilon = U_1 - U_0$. Se supusermos que

$$\begin{pmatrix} U_1 \\ U_0 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma_{11} & \sigma_{10} \\ \sigma_{10} & \sigma_{00} \end{pmatrix} \right]$$

então $U_1 - U_0 \sim N[0, \sigma_{11} + \sigma_{00} - 2\sigma_{10}]$, e nesse caso a normalização $var(\varepsilon) = 1$ equivale a $\sigma_{11} + \sigma_{00} - 2\sigma_{10} = 1$ ou $(\sigma_{11} + \sigma_{00} - 1)/2 = \sigma_{10}$.

(c) Logit: $\Pr(d=1|X) = \Lambda(X\beta) = \frac{\exp(X\beta)}{1+\exp(X\beta)}$. A distribuição de probabilidades de ε que racionaliza o modelo neste caso é a distribuição logística padrão. É possível mostrar que o modelo de utilidades aleatórias com U_1 e U_0 sendo ambos distribuídos segundo uma

distribuição de valor extremo do tipo I, gera uma distribuição de $U_1 - U_0$ logística.

(d) Modelo de complementaridade log-log: $\Pr(d = 1|X) = W(X\beta) = 1 - \exp[-\exp(X\beta)]$. A distribuição de ε neste caso é a de valor extremo do tipo I.

Decidindo que especificação usar Cada especificação tem vantagens e desvantagens.

(a) MPL: a popularidade deste modelo vem da simplicidade de sua estimação, especialmente nos tempos em que os computadores tinham capacidade limitada e estimar modelos probabilísticos usando grandes bases de dados era computacionalmente inviável. Outra vantagem é que o cômputo dos efeitos marginais de mudanças nas covariadas sobre a probabilidade de que $d = 1$ é imediatamente dada pelos respectivos coeficientes β.

O primeiro problema com este modelo é a heterocedasticidade. Para ver isso, note que se $\varepsilon = d - X\beta$, então:

$$\begin{aligned} \varepsilon &= 1 - X\beta \text{ com probabilidade } X\beta \\ &= -X\beta \text{ com probabilidade } 1 - X\beta \end{aligned}$$

E a variância de uma Bernoulli pode ser escrita como $var(z; p) = p(1 - p)$, o que implica:

$$var(\varepsilon|X) = (1 - X\beta) X\beta$$

que claramente depende de X. De cara, este problema descarta a conveniência do uso de MQO para estimar eficientemente os parâmetros β. Pior do que isso: é perfeitamente possível que as probabilidades preditas de que d seja 0 ou 1 para determinados indivíduos se encontre fora do intervalo $[0, 1]$, o que não faz sentido probabilístico e ainda produz variâncias

negativas. Este problema fica ainda mais grave se nosso objetivo for extrapolar as probabilidades estimadas para pontos fora do suporte amostral de (X, d), que é frequentemente o caso quando se trata de simular potenciais efeitos de mudanças nas distribuições das covariadas ou no ambiente econômico sobre a probabilidade de que $d = 1$.

(b) Logit versus Probit: ambas tem formato semelhante (forma de sino, simétrica em torno de zero), com as caudas do Logit sendo ligeiramente mais pesadas. Como pode ser visto no início desta subseção, a distribuição logística possui forma analítica fechada, enquanto a normal acumulada é uma integral que necessita ser computada numericamente. Isso significa que o custo computacional de estimar o Logit é significativamente menor do que o Probit. A vantagem teórica da distribuição Normal é que em muitos casos tem o Teorema do Limite Central para justificar seu uso, enquanto a escolha do Logit pode parecer mais arbitrária.

Além disso, quando o problema envolve mais de duas escolhas é possível mostrar que o Logit sofre do problema de Independência de Alternativas Irrelevantes (Blue Bus, Red Bus), que pode ser enunciado da seguinte forma: suponha que as escolhas relevantes para um agente sejam J, mas o econometrista erroneamente incluiu uma $(J + 1)^a$ alternativa no modelo, completamente redundante a alguma das já incluídas. Idealmente, gostaríamos que os coeficientes que descrevem a probabilidade de que as alternativas não-redundantes que já estavam no modelo não se alterassem com a inclusão da alternativa irrelevante, mas é possivel mostrar inequivocamente que todas as probabilidades das alternativas não-redundantes caem com a inclusão de novas alternativas no modelo. O problema foi mostrado por McFadden (1974), que ilustrou a situação de um indivíduo decidindo se vai de carro ou de ônibus para

o trabalho, sendo completamente indiferente a respeito da cor do ônibus. A especificação correta do problema seria portanto incluir apenas as alternativas carro e ônibus, mas o analista resolveu "caprichar" e colocou as alternativas carro, ônibus azul e ônibus vermelho dentre as possibilidades. Como o indivíduo é indiferente, a probabilidade de ir de carro não deveria mudar com a nova especificação, mas isso infelizmente não ocorre no Logit. A causa dessa limitação é que o Logit multinomial não deixa muita flexibilidade na determinação dos segundos momentos da distribuição (matriz de covariâncias).

O Probit tem como distribuição implícita a Normal padrão, com variância unitária, ao passo que no Logit a distribuição é a logística padrão, com variância igual a $\pi^2/3$.

(c) Distribuição de valores extremos do tipo I: é a única dentre as descritas que assume uma distribuição assimétrica para ε. Seu uso é menos frequente, mas é bastante apropriada para casos em que a ocorrência de $d = 0$ ou de $d = 1$ é rara (tipicamente, os modelos Probit e Logit estimam de forma mais acurada os efeitos marginais e probabilidades preditas quando estas probabilidades estão próximas de 1/2).

2.3.3 Efeitos marginais

Assim como no caso das regressões lineares, estamos interessados no efeito de uma mudança marginal em uma variável explicativa, x_k, sobre nossa variável dependente. O problema nesse caso é que a variável dependente agora é d, que é discreta, e $d(d)/dx_k$ não é definida. Em compensação, a regressão $E(d|X) = \Pr(d = 1|X)$ é uma variável contínua, e podemos calcular o efeito (médio) marginal de x_k sobre d, via $dE(d|X)/dx_k$.

Além disso, ao contrário da regressão linear, o efeito marginal de x_k não é simplesmente

β_k, como antes, pois nosso modelo agora já não é mais linear:

$$\begin{aligned}\frac{dE(d|X)}{dx_k} &= \frac{\partial F(X\beta)}{\partial(X\beta)}\frac{\partial X\beta}{\partial x_k} \\ &= f(X\beta)\beta_k\end{aligned}$$

onde $f(.)$ é a derivada da distribuição acumulada e denota portanto a densidade de ε avaliada no ponto $X\beta$. Vamos então ver como os efeitos marginais são derivados nos diferentes modelos:

(a) MPL: $f(X\beta) = 1$, e portanto o efeito marginal é simplesmente β_k

(b) Probit: $f(X\beta) = \frac{X\beta}{\sqrt{2\pi}}\exp -\frac{(X\beta)^2}{2} = \phi(X\beta)$, e o efeito marginal fica $\phi(X\beta)\beta_k$

(c) Logit: $f(X\beta) = \frac{\exp(X\beta)}{1+\exp(X\beta)}\left(1 - \frac{\exp(X\beta)}{1+\exp(X\beta)}\right) = \Lambda(X\beta)(1-\Lambda(X\beta))$. O efeito marginal é $\Lambda(1-\Lambda)\beta_k$

(d) Log-log: $f(X\beta) = \exp(-\exp(X\beta))\exp(X\beta)$. O efeito marginal é então $[1+W(X\beta)]\exp(X\beta)\beta_k$

Comparando efeitos marginais do Probit com o Logit Considerando as variâncias das respectivas distribuições padrão, o fator aproximado de conversão de coeficientes Logit em coeficientes Probit deveria ser

$$\begin{aligned}\frac{\beta_{Logit}}{\pi^2/3} &\approx \frac{\beta_{probit}}{1} \\ \beta_{Logit} &\approx 1,8\beta_{probit}\end{aligned}$$

No entanto, através de sucessivas tentativas Amemya sugere que, no centro da distribuição, o fator de conversão que melhor aproxima os efeitos marginais nas duas dis-

tribuições é

$$\beta_{Logit} \approx 1,6\beta_{probit}$$

2.3.4 Estimação e teste de hipótese

O método mais popular utilizado para estimar os coeficientes β dos modelos de escolha discreta mencionados anteriormente é o de máxima verossimilhança.

Supondo que nossa base de dados consiste em uma amostra aleatória de $i = 1, ..., N$ indivíduos com informações sobre suas decisões, d_i, e características que potencialmente afetam estas decisões, X_i, podemos escrever a probabilidade conjunta de observarmos $\{d_i, X_i\}_{i=1}^N$ como:

$$\begin{aligned}
\Pr\left(\{d_i, X_i\}_{i=1}^N ; \beta\right) &= \pounds(\beta) \\
&= \prod_{i=1}^N \Pr(d_i, X_i; \beta) \\
&= \prod_{i=1}^N \Pr(d_i | X_i; \beta) \Pr(X_i) \\
&= \prod_{i=1}^N [1 - F(X_i\beta)]^{1-d_i} F(X_i\beta)^{d_i} \Pr(X_i)
\end{aligned}$$

Nosso objetivo então é encontrar o valor de β que maximize a probabilidade de observarmos os dados disponíveis. Sabemos que o valor que maximiza $\Pr\left(\{d_i, X_i\}_{i=1}^N ; \beta\right)$ e $\ln\left[\Pr\left(\{d_i, X_i\}_{i=1}^N ; \beta\right)\right]$ é o mesmo, e portanto podemos escrever nosso problema como:

$$\widehat{\beta}^{MV} = \arg\max_b \{(1 - d_i) \ln[1 - F(X_i\beta)] + d_i \ln[F(X_i\beta)]\}$$

Dentro deste arcabouço, estaremos interessados em examinar os seguintes procedimentos

de estimação e teste:

A. Estimação de $\widehat{\beta}$

B. Estimação da variância assintótica de $\widehat{\beta}$

C. Estimação das probabilidades preditas, $F\left(X\widehat{\beta}\right), 1-F\left(X\widehat{\beta}\right)$, e de suas variâncias assintóticas

D. Estimação dos efeitos marginais, $\frac{d\Pr(d=1|X)}{dX}$, e de suas variâncias assintóticas

E. Testes de hipótese

F. Mensuração do ajuste do modelo aos dados

Estimação de $\widehat{\beta}$ Na maximização do logarítmo da função de verossimilhança acima, precisamos considerar as condições de primeira e segunda ordens (necessárias e suficientes para encontrar o máximo):

i. **Condições de primeira ordem (CPOs)**

$$\begin{aligned}\frac{dF(x\beta)}{\partial\beta} &= \frac{\partial F(x\beta)}{\partial x\beta}\frac{\partial x\beta}{\partial\beta} \\ &= f(x\beta)\,x'\end{aligned}$$

$$\begin{aligned}\frac{d\ln\pounds(\beta)}{d\beta} &= \sum_{i=1}^{N} -(1-d_i)\frac{f(x_i\beta)\,x_i'}{1-F(x_i\beta)} + d_i\frac{f(x_i\beta)\,x_i'}{F(x_i\beta)} \\ &= \sum_{i=1}^{N}\frac{[d_i-F(x_i\beta)]}{[1-F(x_i\beta)]}\frac{f(x_i\beta)}{F(x_i\beta)}x_i' = 0\end{aligned}$$

i.i. *MPL*:

$$\sum_{i=1}^{N} \frac{[d_i - x_i\beta]}{[1 - x_i\beta]} \frac{x_i'}{x_i\beta} = 0$$
$$\sum_{i=1}^{N} \frac{[d_i - x_i\beta]}{var(\varepsilon_i)} x_i' = 0$$
$$\widehat{\beta} = \frac{\sum_{i=1}^{N} \frac{d_i x_i'}{var(\varepsilon_i)}}{\sum_{i=1}^{N} \frac{x_i x_i'}{var(\varepsilon_i)}}$$

bastante semelhante com MQG. Se valesse a homocedasticidade, o estimador se reduziria a MQO.

i.ii. *Probit*: Defina $\Phi_i = F(x_i\beta), \phi_i = f(x_i\beta)$. As CPOs ficam:

$$\sum_{i=1}^{N} \frac{[d_i - \Phi_i]}{[1 - \Phi_i]} \frac{\phi_i}{\Phi_i} x_i' = 0$$

Dado que:

$$\begin{aligned} E(z|z > x\beta) &= -\frac{\phi_i}{1 - \Phi_i} = \lambda_{i0} \\ E(z|z < x\beta) &= \frac{\phi_i}{\Phi_i} = \lambda_{i1} \end{aligned}$$

temos:

$$\sum_{i=1}^{N} \lambda_i x_i = 0$$

onde $\lambda_i = \lambda_{i0}$ se $d_i = 0$ e $\lambda_i = \lambda_{i1}$ se $d_i = 1$.

Neste caso, as CPOs não resultam em um sistema de equações lineares em β, ao contrário do MPL.

i.iii. *Logit*: A densidade assumida do logit é $F(x_i\beta) = \left(1 + e^{-x_i\beta}\right)^{-1} = \Lambda_i$. Logo, as

condições de primeira ordem ficam:

$$\sum_{i=1}^{N} \frac{(d_i - F(x_i\beta))\, f(x_i\beta)}{(1 - F(x_i\beta))\, F(x_i\beta)} x_i = 0$$

$$\Leftrightarrow$$

$$\sum_{i=1}^{N} \frac{(d_i - \Lambda_i)\, \Lambda_i\, (1 - \Lambda_i)}{(1 - \Lambda_i)\, \Lambda_i} x_i = 0$$

$$\Leftrightarrow$$

$$\sum_{i=1}^{N} (d_i - \Lambda_i)\, x_i = 0$$

Curioso notar que $\varepsilon_i = d_i - \Lambda_i$, e que portanto as CPOs do problema ficam $\sum_{i=1}^{N} \varepsilon_i x_i = 0$, semelhantes às condições de momento dos modelos lineares.

ii. Condições de segunda ordem

O propósito desta subseção é apenas mostrar que funções de verossimilhança produzidas por modelos MPL, Probit e Logit são globalmente côncavas, e que portanto as CPOs de fato resultam no estimador desejado. Para ser côncava, a função $\ln[\pounds(\beta)]$ deve ter a seguinte propriedade:

$$\frac{\partial^2 \ln[\pounds(\beta)]}{\partial\beta\partial\beta'} < 0$$

Sabemos que:

$$\begin{aligned}
\frac{\partial^2 \ln[\pounds(\beta)]}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial(x\beta)} \left[\frac{\partial \ln[\pounds(\beta)]}{\partial(x\beta)} x\right] \frac{\partial(x\beta)}{\partial\beta} \\
&= \frac{\partial^2 \ln[\pounds(\beta)]}{\partial(x\beta)\,\partial(x\beta)'} xx' \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial(x_i\beta)} \frac{[d_i - F(x_i\beta)]}{[1 - F(x_i\beta)]} \frac{f(x_i\beta)}{F(x_i\beta)} x_i x_i'
\end{aligned}$$

ii.i. *MPL*:

$$\sum_{i=1}^{N} \frac{\partial}{\partial(x_i\beta)} \left[\frac{[d_i - x_i\beta]}{[1-x_i\beta]} \frac{1}{x_i\beta} x_i\right] x_i'$$
$$= \sum_{i=1}^{N} - \left[\frac{1}{[1-x_i\beta]} \frac{x_i}{x_i\beta} + \frac{[d_i - x_i\beta]x_i}{[1-x_i\beta]^2 (x_i\beta)^2} (1 - 2x_i\beta)\right] x_i'$$
$$= \sum_{i=1}^{N} - \left[\frac{d_i - x_i\beta}{[1-x_i\beta]x_i\beta}\right]^2 x_i x_i' < 0$$

(usando o fato de que $d_i^2 = d_i$, dado que $d_i \in \{0, 1\}$).

ii.ii *Probit*: aqui vamos usar o seguinte fato: $\phi'(z) = -z\phi(z)$. Precisamos mostrar então que:

$$\sum_{i=1}^{N} \frac{\partial}{\partial(x_i\beta)} (\lambda_i x_i) x_i' = \sum_{i=1}^{N} \frac{\partial \lambda_i}{\partial(x_i\beta)} x_i x_i' < 0$$

no caso de $\lambda_i = \lambda_{0i} = -\frac{\phi(x_i\beta)}{1-\Phi(x_i\beta)}$:

$$\begin{aligned}
\frac{\partial \lambda_i}{\partial(x_i\beta)} &= \frac{\partial}{\partial(x_i\beta)} \left[-\frac{\phi(x_i\beta)}{1-\Phi(x_i\beta)}\right] \\
&= \frac{x_i\beta\phi(x_i\beta)}{1-\Phi(x_i\beta)} - \left(\frac{\phi(x_i\beta)}{1-\Phi(x_i\beta)}\right)^2 \\
&= -\lambda_{0i}[x_i\beta + \lambda_{0i}] < 0
\end{aligned}$$

e no caso de $\lambda_i = \lambda_{1i} = \frac{\phi(x_i\beta)}{\Phi(x_i\beta)}$:

$$\begin{aligned}
\frac{\partial \lambda_i}{\partial(x_i\beta)} &= \frac{\partial}{\partial(x_i\beta)} \left[\frac{\phi(x_i\beta)}{\Phi(x_i\beta)}\right] \\
&= -\left[\frac{x_i\beta\phi(x_i\beta)}{\Phi(x_i\beta)} + \left(\frac{\phi(x_i\beta)}{\Phi(x_i\beta)}\right)^2\right] \\
&= -\lambda_{1i}[x_i\beta + \lambda_{1i}] < 0
\end{aligned}$$

e a CSO fica:

$$-\sum_{i=1}^{N} \lambda_i [x_i\beta + \lambda_i] < 0$$

ii.iii *Logit*: derivando a CPO com respeito a β, temos:

$$\begin{aligned} & \frac{\partial}{\partial (x_i\beta)} \left[\sum_{i=1}^{N} (d_i - \Lambda_i) x_i \right] x_i' \\ = & -\sum_{i=1}^{N} \Lambda_i (1 - \Lambda_i) x_i x_i' < 0 \end{aligned}$$

Como todos os modelos produzem funções de (log) verossimilhança globalmente côncavas, o método de Newton-Raphson deve encontrar o máximo e poucas iterações (convergência é relativamente rápida).

Estimação de Avar$\left(\widehat{\beta}\right)$ Após estimarmos os parâmetros do modelo, $\widehat{\beta}$, o passo seguinte é testar sua significância estatística. Tipicamente, a distribuição de $\widehat{\beta}$ numa amostra pequena é difícil de ser tratada tanto analitica quanto numericamente, e os testes recorrem então às propriedades assintóticas dos estimadores. Aplicando o método delta e teoremas de Slutsky, temos que:

$$\begin{aligned} & \sqrt{N}\left(\widehat{\beta} - \beta^*\right) \xrightarrow{d} \\ & N\left[0, E\left(-\frac{\partial^2 \pounds(\beta^*)}{\partial\beta\partial\beta'}\right)^{-1} \left(E\frac{\partial \pounds(\beta^*)}{\partial\beta} E\frac{\partial \pounds(\beta^*)}{\partial\beta'}\right) E\left(-\frac{\partial^2 \pounds(\beta^*)}{\partial\beta\partial\beta'}\right)^{-1}\right] \end{aligned}$$

Sob condições de regularidade, satisfeitas pelos modelos aqui tratados, pode-se mostrar que:

$$E\left(\frac{\partial^2 \pounds(\beta^*)}{\partial\beta\partial\beta'}\right) = E\frac{\partial \pounds(\beta^*)}{\partial\beta} E\frac{\partial \pounds(\beta^*)}{\partial\beta'}$$

e que portanto:

$$\sqrt{N}\left(\widehat{\beta}-\beta^{*}\right)\overset{d}{\rightarrow}N\left[0,-I\left(\beta^{*}\right)^{-1}\right]$$

onde $-I\left(\beta^{*}\right)^{-1}$ atinge o limite inferior de Cramer-Rao e é definido como:

$$I\left(\beta^{*}\right)=p\lim_{N\rightarrow\infty}\left(\frac{1}{N}\frac{\partial^{2}\pounds\left(\beta^{*}\right)}{\partial\beta\partial\beta'}\right)$$

A variância assintótica de $\widehat{\beta}$ é, portanto:

$$Avar\left(\widehat{\beta}\right)=-\left[NI\left(\beta^{*}\right)\right]^{-1}$$

Os resultados acima sugerem 2 diferentes formas de representar $Avar^{-1}\left(\widehat{\beta}\right)$:

$$-E\left(\frac{\partial^{2}\pounds\left(\beta^{*}\right)}{\partial\beta\partial\beta'}\right)$$
$$E\frac{\partial\pounds\left(\beta^{*}\right)}{\partial\beta}E\frac{\partial\pounds\left(\beta^{*}\right)}{\partial\beta'}$$

e cada uma deles inspira diferente estimadores, todos baseados na convergência de $g\left(\widehat{\beta}\right)$ para $g\left(\beta^{*}\right)$, para funções contínuas e diferenciáveis de β, conforme a amostra cresce de N para infinito.

O primeiro destes estimadores é

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^{2}}{\partial\beta\partial\beta'}\left[\left(1-d_{i}\right)\ln\left[1-F\left(X_{i}\beta\right)\right]+d_{i}\ln\left[F\left(X_{i}\beta\right)\right]\right]$$

ou seja, o negativo da média das contribuições individuais para o Hessiano da função de log-verossimilhança, e possui a desvantagem de poder eventualmente não ser positivo-definido (ainda que assintoticamente deva convergir para uma matriz que o seja).

O segundo parte da definição de vetor de score dos indivíduos,

$$s_i(\beta) = \frac{\partial}{\partial \beta}\left[(1-d_i)\ln\left[1-F(x_i\beta)\right] + d_i \ln\left[F(x_i\beta)\right]\right]$$

para definir o estimador:

$$\frac{1}{N}\sum_{i=1}^{N} s_i(\beta)\left[s_i(\beta)\right]'$$

Este estimador foi proposto por Berndt, Hall, Hall e Hausman (1974), e é chamado de BHHH, mas possui a desvantagem de não ser bem comportado em amostras pequenas.

Finalmente, o terceiro estimador bastante usado na literatura é:

$$-\frac{\partial^2}{\partial\beta\partial\beta'}\left[\frac{1}{N}\sum_{i=1}^{N}(1-d_i)\ln\left[1-F(X_i\beta)\right] + d_i\ln\left[F(X_i\beta)\right]\right]$$

O problema com esta alternativa é que nem sempre há uma forma fechada para esta expressão, e mesmo seu cômputo numérico pode ser difícil. Se isso não for um problema, como nos casos tratados nestas notas, este estimador possui as vantagens de (i) envolver apenas primeiras derivadas de médias e variâncias condicionais, (ii) quando existir, precisa ser positivo-definido, e (iii) possui melhores propriedades em pequenas amostras que os demais.

Estimação de Avar$\left[F\left(X\widehat{\beta}\right)\right]$ Pelo método Delta, sabemos que:

$$\sqrt{N}\left(\widehat{\theta}-\theta^*\right) \xrightarrow{d} N(0,V)$$

então para qualquer função contínua e diferenciável do estimador, $g(\theta)$, teremos que:

$$\sqrt{N}\left(g\left(\widehat{\theta}\right)-g\left(\theta^{*}\right)\right)\xrightarrow{d}N\left(0,g'\left(\theta^{*}\right)^{T}Vg'\left(\theta^{*}\right)\right)$$

onde $g'\left(\theta^{*}\right)=\partial g\left(\theta^{*}\right)/\partial\theta$, e o superscrito T denota transposição.

No caso particular de $\widehat{P}=F\left(X\widehat{\beta}\right)$, este resultado sugere que:

$$\sqrt{N}\left(F\left(X\widehat{\beta}\right)-F\left(X\beta^{*}\right)\right)\xrightarrow{d}N\left(0,-\left[xf\left(X\beta^{*}\right)\right]^{T}I\left(\beta^{*}\right)^{-1}\left[xf\left(X\beta^{*}\right)\right]\right)$$

e como $f\left(X\beta^{*}\right)$ é escalar:

$$\sqrt{N}\left(F\left(X\widehat{\beta}\right)-F\left(X\beta^{*}\right)\right)\xrightarrow{d}N\left(0,-f\left(X\beta^{*}\right)^{2}x^{T}I\left(\beta^{*}\right)^{-1}x\right)$$

Um estimador natural de $avar\left[F\left(X\widehat{\beta}\right)\right]$ é portanto:

$$avar\left[F\left(X\widehat{\beta}\right)\right]=f\left(X\widehat{\beta}\right)^{2}x^{T}Avar\left(\widehat{\beta}\right)x$$

Note que podemos (e quase sempre temos que) estimar $avar\left[F\left(X\widehat{\beta}\right)\right]$ para cada vetor $X=x_i$, que diferentemente de $Avar\left(\widehat{\beta}\right)$, varia com X.

Estimação de Avar$\left[f\left(X\widehat{\beta}\right)\widehat{\beta}\right]$ Como vimos, $f\left(X\widehat{\beta}\right)\widehat{\beta}$ mede o efeito marginal de uma mudança em X sobre a probabilidade de que $d=1$. Assim como no exemplo acima, esta também será na maioria das vezes uma função contínua e diferenciável de β, e para a qual o método Delta se aplica. Definindo $\widehat{\gamma}=f\left(X\widehat{\beta}\right)\widehat{\beta}$, temos que:

$$\begin{aligned} Avar\left(\widehat{\gamma}\right) &= \left(\frac{\partial\widehat{\gamma}}{\partial\beta}\right)^{T} avar\left(\widehat{\beta}\right)\left(\frac{\partial\widehat{\gamma}}{\partial\beta}\right) \\ \frac{\partial\widehat{\gamma}}{\partial\beta} &= \frac{\partial\left[f\left(X\widehat{\beta}\right)\widehat{\beta}\right]}{\partial\beta} \\ &= f\left(X\widehat{\beta}\right)\frac{\partial\widehat{\beta}}{\partial\beta} + \widehat{\beta}X^{T}\frac{\partial f\left(X\widehat{\beta}\right)}{\partial X\beta} \\ &= If\left(X\widehat{\beta}\right) + \widehat{\beta}X^{T}\frac{\partial f\left(X\widehat{\beta}\right)}{\partial X\beta} \end{aligned}$$

Vejamos como ficam essas expressões nos modelos probabilísticos tratados:

i. *MPL*: Neste modelo, $f(z) = 1$ e $f'(z) = 0$, de modo que $Avar\left(\widehat{\gamma}\right) = avar\left(\widehat{\beta}\right)$

ii. *Probit*: $f(z) = \phi(z)$, $f'(z) = -z\phi(z)$, onde $\phi(z) = (2\pi)^{-1/2}\exp\left(-z^2/2\right)$ é a densidade da distribuição Normal padrão.

$$Avar\left(\widehat{\gamma}\right) = \left[\phi\left(x\widehat{\beta}\right)\right]^{2}\left(I - \left(x\widehat{\beta}\right)\widehat{\beta}x^{T}\right)avar\left(\widehat{\beta}\right)\left(I - \left(x\widehat{\beta}\right)\widehat{x\beta}^{T}\right)$$

iii. *Logit*: $f(z) = \Lambda(1-\Lambda)$, $f'(z) = \Lambda(1-\Lambda)(1-2\Lambda)$.

$$Avar\left(\widehat{\gamma}\right) = \left[\widehat{\Lambda}\left(1-\widehat{\Lambda}\right)\right]^{2}\left(I - \left(1-2\widehat{\Lambda}\right)\widehat{\beta}x^{T}\right)avar\left(\widehat{\beta}\right)\left(I - \left(1-2\widehat{\Lambda}\right)\widehat{x\beta}^{T}\right)$$

Testes de hipótese Na maioria dos trabalhos empíricos, os autores estão interessados em testar uma restrição linear sobre os parâmetros, do tipo $H\beta = h$, onde o rank p da matriz H mede o número de restrições que (conjuntamente) queremos testar. Note que $\beta_k = 0$ para algum k específico é apenas um caso particular do tratado acima. Há três formas de testar $H\beta = h$ que predominam na literatura:

Teste de Wald Defina $\gamma(\beta) = H\beta - h$. Então:

$$\sqrt{N}\left[\gamma\left(\widehat{\beta}\right) - \gamma(\beta^*)\right] \xrightarrow{d} N\left(0, NH^T avar(\beta^*) H\right)$$

e sob a hipótese nula, $\gamma(\beta^*) = 0$. Portanto:

$$\gamma\left(\widehat{\beta}\right)^T \left[H^T avar(\beta^*) H\right]^{-1} \gamma\left(\widehat{\beta}\right) \overset{A}{\sim} \chi^2(p)$$

Teste de razão de verossimilhança Suponha que estimemos o modelo duas vezes: uma irrestrita e outra impondo $\gamma(\beta) = 0$. Então, pode-se mostrar que:

$$-2\ln\left[\frac{\pounds\left(\widehat{\beta}^R\right)}{\pounds\left(\widehat{\beta}\right)}\right] \overset{A}{\sim} \chi^2(p)$$

onde $\pounds\left(\widehat{\beta}^R\right)$ denota a função de verossimilhança avaliada nos coeficientes estimados do modelo restrito.

Teste do multiplicador de Lagrange (ou teste de Score) Uma forma de escrever o problema de maximizar o logarítmo da função de verossimilhança sujeito à restrição $\gamma(\beta^*) = 0$ é escrever o problema como:

$$\max_b \left[\ln \pounds(\beta) - \lambda\left(\beta - \beta^R\right)\right]$$

Neste caso, a condição de primeira ordem é:

$$\frac{\partial \ln \pounds(\beta)}{\partial \beta} = \lambda$$

e a restrição só será efetiva se $\lambda > 0$ (isto é, $H_0 : \lambda = 0$). Ora, como λ é em si uma função de β, teremos que:

$$\lambda\left(\widehat{\beta}^R\right)^T avar\left(\widehat{\beta}\right)\lambda\left(\widehat{\beta}^R\right) \overset{A}{\sim} \chi^2(p)$$

onde $\lambda\left(\widehat{\beta}^R\right)$ é a funçãoλ avaliada no valor restrito dos coeficientes. Este teste é usado principalmente em situações onde a estimação do modelo irrestrito é difícil.

Teste de restrições não-lineares Uma variante do teste de Wald disponível para restrições genéricas do tipo $H(\beta) = 0$ é aplicar diretamente o método delta e obter a estatística:

$$W = H\left(\widehat{\beta}\right)^T \left[\nabla_\beta H\left(\widehat{\beta}\right) avar\left(\widehat{\beta}\right) \nabla_\beta H\left(\widehat{\beta}\right)^T\right]^{-1} H\left(\widehat{\beta}\right)$$

onde $\nabla_\beta H\left(\widehat{\beta}\right)$ denota o Jacobiano da função $H(.)$ com respeito a β avaliado em $\widehat{\beta}$. O método sugere que:

$$W \sim \chi^2_q$$

onde q é o rank da matriz jacobiana de $H(.)$ avaliada em $\widehat{\beta}$.

Medindo o ajuste do modelo aos dados Há três medidas básicas do ajuste de um modelo com variável dependente discreta aos dados:

Logarítmo da função de verossimilhança avaliada no parâmetro estimado Comparação direta do logarítmo da função de verossimilhança em um modelo estimado onde o único regressor é uma constante (intercepto), com o logarítmo desta mesma função quando todos os regressores são incluídos. Mede a contribuição da inclusão dos regressores para a função de verossimilhança (e portanto o aumento na probabilidade de que os dados tenham vindo de um mundo relativamente bem descrito pelo modelo versus um mundo sem modelo algum).

Índice da razão de verossimilhança Guarda similaridade com o R^2 em uma regressão linear, e é definido como $1 - \frac{\ln \pounds\left(\widehat{\beta}\right)}{\ln \pounds(0)}$. Se o modelo tem um ajuste perfeito, a probabilidade de que os dados tenham vindo de um mundo descrito por ele é $\pounds\left(\widehat{\beta}\right) = 1$, o que implica que o indicador fica $1 - 0 = 1$. Por outro lado, se o modelo explica muito pouco do mundo, a inclusão das covariadas muda pouco a função de verossimilhança, e o indicador se aproxima de 0. No Stata, essa estatística é chamada de pseudo-R^2 (e em alguns livros, de R^2 de McFadden). Devemos ter cuidado, contudo, pois parte da literatura usa outra definição para pseudo R^2 (ou R^2 de Amemya), inspirada no mesmo princípio mas com fórmula distinta: $1 - \frac{1}{1+2\left(\ln \pounds\left(\widehat{\beta}\right) - \ln \pounds(0)\right)/N}$.

Tabela de acertos e erros Para cada indivíduo, defina como $\widehat{d_i} = 1\left[F\left(x_i\widehat{\beta}\right) > p\right]$, para algum valor p definido pelo econometrista (usualmente 1/2). Com isso queremos dizer que, para um indivíduo com características x_i, a probabilidade predita de que tome a decisão $d = 1$ é $F\left(x_i\widehat{\beta}\right)$. Se esta probabilidade ultrapassar um certo nível p, dizemos que a ação

predita para ele pelo modelo é $\widehat{d}_i$. Na tabela, comparamos as decisões preditas com as efetivamente tomadas para julgar quão bem o modelo se ajusta aos dados. O problema com essa estratégia é que o valor p é arbitrário. Mais detalhes em Greene, pg.652.

2.4 Escolhas simultâneas

Ainda que tenha sido mencionado que o modelo de utilidades aleatórias pode em princípio motivar a análise de um problema com múltiplas escolhas disponíveis, quase toda a nossa análise esteve concentrada em situações onde os agentes decidem entre duas opções. Dentre as principais características do modelo, vimos que não é possível identificar simultaneamente a função-resposta $H_{jk}(X) = v_j(X) - v_k(X)$, que associa características observáveis à recompensa associada à decisão j, e a distribuição dos resíduos, $F_\varepsilon(c; X, d) = \Pr(\varepsilon < c|X, d)$, e tipicamente optamos por assumir uma forma funcional específica para essa distribuição[4].

Com múltiplas escolhas, nem sempre é fácil determinar que objetos conseguimos identificar, ainda que nos modelos mais conhecidos essa questão já tenha sido respondida. Além disso, veremos que a estimação do modelo pode simplificar enormemente se pudermos supor que os elementos não observáveis associados à recompensa de cada alternativa são independentes entre si. Finalmente, maior atenção será dada ao Logit multinomial, por ser o modelo mais utilizado na literatura, e a comparações deste com outros modelos.

[4] A rigor, há situações em que se pode identificar parte da distribuição de ε sem abrir mão da identificação de $H_{jk}(X_i)$, mas não é o que ocorre na maioria dos modelos convencionais. Em particular, se houver um regressor $X_k \in X$ tal que (i) o suporte de X_k contém o suporte de ε, e (ii) a função H_{jk} é do tipo $H_{jk}(X) = h_{jk}(X_{-k})X_k$ ou $H_{jk}(X) = h_{jk}(X_{-k}) + X_k$, onde X_{-k} denota o vetor X sem a k-ésima coordenada, então podemos identificar simultaneamente h_{jk} e a distribuição de ε.

2.4.1 Elementos do modelo

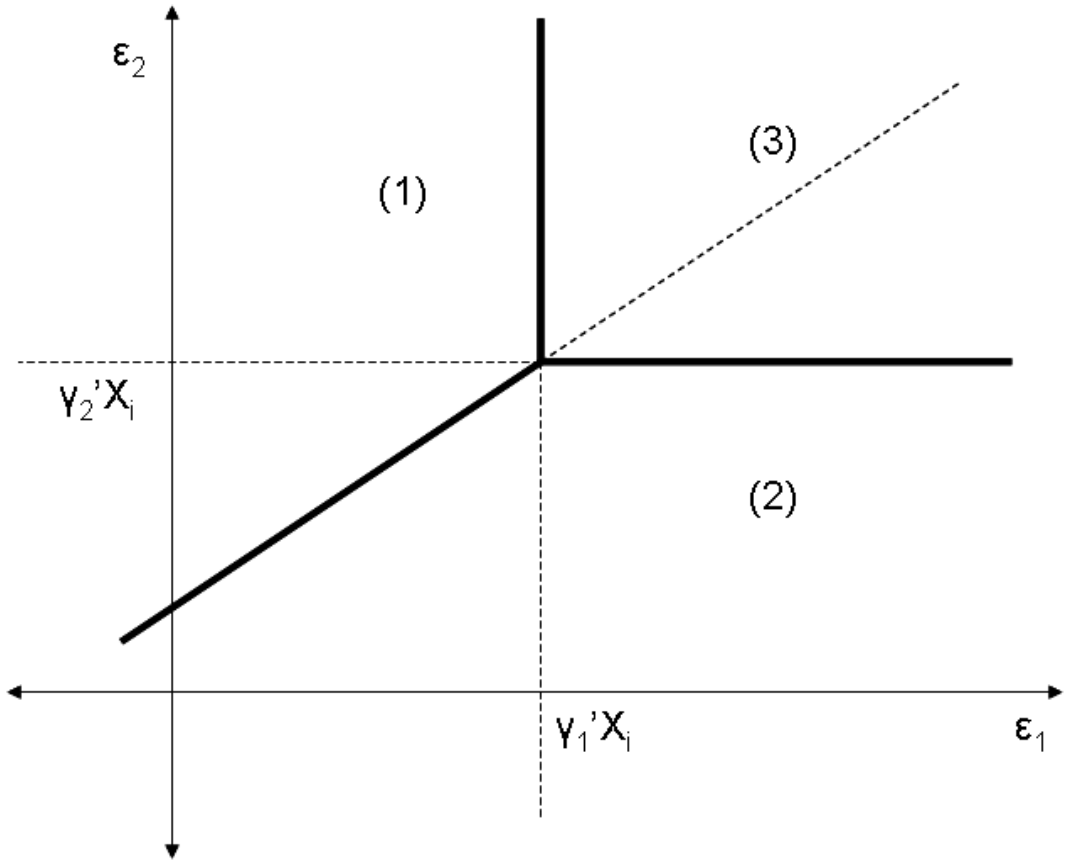
- J alternativas, cada uma com recompensas $v_{ij} + U_{ij} = v_j(X_i) + U_{ij} = \beta_j' X_i + U_{ij}$, onde U_{ij} não é observado pelo econometrista.

- $J-1$ variáveis dummy representando as decisões dos agentes. Definindo uma escolha como padrão (digamos, a escolha J), temos que

$$d_{ij} = 1 \Leftrightarrow \left(\beta_j - \beta_k\right)' X_i > U_{ik} - U_{ij}; \forall k \neq j$$

o que significa que novamente não podemos em geral estimar separadamente β_j, β_k, mas apenas $\gamma_{jk} = \left(\beta_j - \beta_k\right)$. Na verdade, tendo a alternativa J como referência, podemos sem perda de generalidade dizer que os $J-1$ vetores $\gamma_j = \left(\beta_j - \beta_J\right)$ são identificados (basta ver que se γ_{jk} e γ_{kJ} são identificados, então $\gamma_j = \gamma_{jk} - \gamma_{kJ}$ também o é, e que qualquer γ_{jk} pode ser recuperado com informações de γ_j, γ_k, fazendo $\gamma_{jk} = \gamma_j - \gamma_k$). Do mesmo modo, para J alternativas precisamos nos preocupar apenas com a distribuição do vetor de não-observáveis de dimensão $J-1$, $\varepsilon = [\varepsilon_j]$, onde $\varepsilon_j = U_J - U_j$. Com isso temos o resultado geral de que para cada J possibilidades, a dimensão de nosso modelo é sempre uma dimensão menor, e isso se deve ao fato de que decisões são sempre tomadas de modo relativo (ou em outro contexto, que escolhas econômicas são tomadas sempre por comparações, e exatamente por esse motivo transformações monotônicas de preferências são inócuas do ponto de vista decisório).

- Dados: tipicamente, o econometrista tem acesso a características dos agentes, X_i, e decisões, d_{ij}. Em princípio, poderíamos dizer que também características das escolhas podem ser observadas, mas examinaremos esta variante do modelo quando tratarmos dos modelos de escolha condicional.



- Conteúdo empírico do modelo: $E\left(d_j|X_i\right) = \Pr\left(\varepsilon_1 < \varepsilon_j + \left(\gamma_j - \gamma_1\right)' X_i, \varepsilon_2 < \varepsilon_j + \left(\gamma_j - \gamma_2\right)' X_i, ..., \varepsilon_{J-}\right.$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\varepsilon_j + \left(\gamma_j - \gamma_1\right)' X_i} ... \int_{-\infty}^{\varepsilon_j + \left(\gamma_j - \gamma_{J-1}\right)' X_i} f\left(\varepsilon_1, \varepsilon_2, ..., \varepsilon_j, ..., \varepsilon_{J-1}\right) d\varepsilon_1 d\varepsilon_2, ..., \varepsilon_j, ...d\varepsilon_{J-1}$$

$$= \int_{-\infty}^{\infty} \frac{\partial F}{\partial \varepsilon_j}\left(\varepsilon_j + \left(\gamma_j - \gamma_1\right)' X_i, \varepsilon_j + \left(\gamma_j - \gamma_2\right)' X_i, ..., \varepsilon_j + \left(\gamma_j - \gamma_{J-1}\right)' X_i\right) d\varepsilon_j$$

- A figura abaixo ajuda a visualizar as regiões de integração para as probabilidades de cada escolha no caso $J = 3$:

- Caso particular: $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_{J-1})$ independentes (e eventualmente identicamente distribuí-

dos):

$$\Pr\left[\max_j \left\{\varepsilon_j < \gamma_j' X_i\right\}\right] = \prod_{j=1}^{J-1} \Pr\left(\varepsilon_j < \gamma_j' X_i\right)$$

2.4.2 O que podemos identificar com diferentes tipos de dados?

- Proporções agregadas e market share (proporções de agentes que escolhem uma dada alternativa)

- Proporções agregadas em subamostras definidas por categorias das varáveis observáveis (homens, pobres, moradores de Brasília, etc.).

- Probabilidades preditas de que um indivíduo com determinado conjunto de características faça uma dada escolha

- Variações nessas mesmas probabilidades num cenário contrafactual onde as características do indivíduo sejam exogenamente alteradas

- Efeitos marginais

- Alguns parâmetros (estruturais) do modelo: $\gamma' s$ (mas não $\beta' s$. De outro modo, se for plausível normalizar os coeficientes associados à recompensa de uma dada ação, por exemplo fazendo $\beta_0 = 0$, então os demais $\beta' s$ seriam identificados); parte da matriz de covariâncias

$$\underset{(U's)}{\text{Modelo estrutural:}} \quad \Sigma_U = \begin{matrix} \sigma_0^2 & \sigma_{01} & \cdots & \sigma_{0J} \\ \sigma_{01} & \sigma_1^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{0J} & \cdots & \cdots & \sigma_J^2 \end{matrix}$$

$$\underset{(\varepsilon_j = U_0 - U_j)}{\text{somente comparações:}} \quad \Sigma_\varepsilon = \begin{matrix} \widetilde{\sigma}_0^2 & \widetilde{\sigma}_{01} & \cdots & \widetilde{\sigma}_{0J} \\ \widetilde{\sigma}_{01} & \widetilde{\sigma}_1^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \widetilde{\sigma}_{0J} & \cdots & \cdots & \widetilde{\sigma}_J^2 \end{matrix}$$

$$\widetilde{\sigma}_j^2 = \sigma_j^2 + \sigma_0^2 - 2\sigma_{0j}$$

$$\widetilde{\sigma}_{jk} = \sigma_{jk} + \sigma_0^2 - \sigma_{0j} - \sigma_{0k}$$

No conjunto de desigualdades que caracterizam as escolhas, $d_j = 1 \Leftrightarrow$:

$$\begin{aligned} \varepsilon_j &< \gamma_j' X \\ \varepsilon_j - \varepsilon_1 &< \left(\gamma_j - \gamma_1\right)' X \\ &\vdots \\ \varepsilon_j - \varepsilon_J &< \left(\gamma_j - \gamma_J\right)' X \end{aligned}$$

Claramente, se dividirmos todas as desigualdades por uma constante positiva c, nenhuma delas se altera, de modo que observacionalmente ε/c e ε são indistinguíveis e a normalização de uma das entradas de Σ_ε se faz necessária. Note contudo que o fato de que as demais coordenadas de Σ_ε sejam identificadas ainda não implica que possamos determinar unicamente as coordenadas de Σ_U, pois estas envolvem mais parâmetros do que Σ_ε. Tipicamente, depois da normalização inicial sobramos com um sistema de $J\left[(J-1)/2+1\right]$ equações para $(J+1)\left[J/2+1\right]$ incógnitas, de forma que ao menos $(J+1)\left[J/2+1\right] - J\left[(J-1)/2+1\right] =$

$J+1$ normalizações de coordenadas de Σ_U são necessárias para recuperar os demais parâmetros originais. Em geral, opta-se por fixar a diagonal principal de Σ_U em 1.

2.5 Logit multinomial

No caso da distribuição logística multivariada, a distribuição padrão supõe independência entre as coordenadas, o que implicará em propriedades não desejáveis para os resultados do modelo (ver discussão abaixo sobre a independência de alternativas irrelevantes). Não obstante, este é o modelo mais usado para estimar decisões multiplas simultâneas, especialmente devido à simplicidade computacional (no caso geral, e no probit multivariado em particular, a integral múltipla descrita acima não apresenta solução analítica ou "fechada", e deve ser calculada numericamente, o que costuma ser computacionalmente pesado).

No logit, cada coordenada do vetor de não observáveis do modelo de utilidades aleatórias é suposto seguir uma distribuição de valores extremos do tipo I: $f(U_j) = e^{-U_j} e^{-e^{-U_j}}$. A média é $E(U_j) = .5772$ (constante de Euler), e a variância é $var(U_j) = \pi^2/6$. Devido à independência entre as coordenadas, a integral múltipla mostrada acima simplifica para:

$$\begin{aligned} p_j(X_i) &= \Pr[d_{ij} = 1 | X_i] \\ &= \int_{-\infty}^{\infty} e^{-U_{ij}} e^{-e^{-U_{ij}}} \prod_{k \neq j} e^{e^{-\left(\beta_k - \beta_j\right)' X_i - U_{ij}}} dU_{ij} \\ &= \int_{-\infty}^{\infty} e^{-U_{ij}} e^{-e^{-U_{ij}} \left[1 + \Sigma_{k \neq j} e^{\left(\beta_j - \beta_k\right)' X_i}\right]} dU_{ij} \end{aligned}$$

e se definirmos $\lambda_{ij} = \ln\left[1 + \Sigma_{k \neq j} e^{\left(\beta_j - \beta_k\right)' X_i}\right]$:

$$\begin{aligned} p_j(X_i) &= \int_{-\infty}^{\infty} e^{-U_{ij}} e^{-e^{-\left(U_{ij}-\lambda_{ij}\right)}} dU_{ij} \\ &= e^{-\lambda_{ij}} \\ &= \frac{1}{1+\Sigma_{k\neq j} e^{\left(\beta_j-\beta_k\right)' X_i}} \\ &= \frac{e^{\beta_j' X_i}}{1+\Sigma_{k=1}^{J} e^{\beta_k' X_i}} \end{aligned}$$

Claramente, a função de verossimilhança fica:

$$\begin{aligned} \pounds(\beta) &= \prod_{i=1}^{N}\prod_{j=1}^{J}\left[p_j(X_i)\right]^{d_{ij}} \\ \ln \pounds(\beta) &= \sum_{i=1}^{N}\sum_{j=1}^{J} d_{ij} \ln\left[p_j(X_i)\right] \end{aligned}$$

Note que a função sofre apenas alterações menores se estivermos em uma situação onde o conjunto de escolha difere entre agentes (isto é, supondo que alguns decidem entre J opções, outros entre L opções e assim por diante). Em particular, se N_1 agentes possuem J_1 escolhas, N_2 agentes possuem J_2 escolhas, e assim por diante até completar os N_M tipos de agentes em nossa amostra, teremos

$$\pounds(\beta) = \prod_{i=1}^{N_1}\left[p_{j_1}(X_i)\right]^{d_{ij_1}} * ... * \prod_{i=1}^{N_M}\left[p_{jM}(X_i)\right]^{d_{ijM}}$$

$$N_1 + N_2 + ... + N_M = N$$

2.5.1 Propriedades

- Forma fechada (distribuição acumulada não envolve integrais)

- Sem perda de generalidade, podemos rescrever o modelo como $p_j(X_i) = \frac{e^{\frac{\beta_j' X_i}{\sigma}}}{1+\Sigma_{k=1}^{J} e^{\frac{\beta_k' X_i}{\sigma}}}$. Assim como no caso binário, o reescalonamento não é identificado e leva às mesmas relações observadas nos dados. No entanto:

  $\lim_{\sigma \to \infty} p_j(X_i) = 1/J$, sugerindo que as probabilidades de escolha passam a independer de X_i

  $\lim_{\sigma \to 0} p_j(X_i) = 1$, se e somente se $\beta_j' X_i > \beta_k' X_i; \forall k \neq j$. Probabilidades degeneradas

- Independência de Alternativas Irrelevantes (IIA): $\frac{p_j(X_i)}{p_k(X_i)} = e^{-\left(\beta_j - \beta_k\right)' X_i}$, o que independe das utilidades médias ($\beta_h' X_i$) das demais alternativas $h \neq j, k$. Via de regra, essa propriedade limita a substitutibilidade entre alternativas, se uma alternativa se tornar menos atrativa (por exemplo, devido a um aumento de preços ou introdução de novas opções no mercado). Em termos dos parâmetros que poderíamos identificar nos dados, a hipótese de independência entre os não-observáveis impõe que a matriz de covariâncias seja completamente determinada com elementos diagonais igual a $\pi^2/3$, e os demais fixados em zero, jogando fora parte da informação presente nos dados. Por outro lado, há na literatura generalizações da distribuição logística multivariada que tentam incorporar correlações não-triviais entre as coordenadas, mas em geral o que ocorre é que nesses casos a resultante deixa de ter a simplicidade e a solução analítica do logit.

Example 1 *Suponha que o indivíduo deseje comprar um carro e que as opções sejam sedan = 0 ou esportivo verde= 1, ambos com a mesma utilidade média, que por conveniência chamaremos de $-v_j$, isto é, $\beta_1' X = -v_1 = -v_0 = \beta_0' X$. Desse modo:*

$$p_0(v) = \frac{e^{-v_0}}{e^{-v_0} + e^{-v_1}} = \frac{1}{2} = p_1(v)$$

Qual a interpretação de v nesse caso? Em termos da decisão de comprar o carro sedan e transformando utilidade em reais, v_1 representa o preço de reserva que o vendedor pode cobrar do sedan de modo que este ainda seja vendido[5]*, e v_0 o preço efetivamente cobrado. Na decisão das famílias, se $-v_0 > -v_1$ então o sedan é o escolhido, e em caso contrário, não.*

Um novo tipo de carro, esportivo vermelho, é então introduzido na economia. Supondo que o agente seja indiferente com respeito à cor do automóvel, devemos esperar que $v_1 = v_2$, e nesse caso $p_0(v) = p_1(v) = p_2(v) = \frac{e^{-v_2}}{e^{-v_0}+e^{-v_1}+e^{-v_2}} = 1/3$.

Conclusão: o market share (fração de consumidores que escolhem uma dada alternativa) do sedan caiu de 1/2 para 1/3, com a introdução de uma opção que somente deveria competir pela fatia esportiva do mercado. A razão é que, ainda que esportivos verdes e vermelhos tenham mesma utilidade média, o componente não-observado é suposto ser independente entre eles.

Suponha agora que X_i sejam as características do automóvel e que no modelo teórico β_j sejam os preços das características que compõem o carro que os consumidores estariam dispostos a pagar, de modo que o preço máximo que uma firma pode vender o carro (e ainda encontrar compradores) seja v_j. Computando as elasticidades, temos que:

$$\begin{aligned} \varepsilon_{p_j,v_j} &= -(1-p_j(v)) \\ \varepsilon_{p_j,v_k} &= p_k(v) \end{aligned}$$

ou seja, a elasticidade (p_j, v_k) (que pode ser interpretada como elasticidade-preço) só depende do market share da alternativa cujo v_k é alterado, independentemente do nível em que se encontram os demais v_l. Suponha agora que o autor do estudo seja uma firma monopolista de carros desejando estabelecer o preço ótimo dos seus produtos. Se o número de consumidores é m, então a demanda pelo carro do tipo j é $mp_j(v)$. Se o custo de produção do carro for constante, c_j, então o lucro obtido com uma unidade adicional de carro produzido do tipo j é:

$$\pi_j = (v_j - c_j)\, mp_j(v)$$

o problema do monopolista é precificar o bem de modo a maximizar lucro. As condições de primeira ordem nesse caso ficam:

$$v_j - c_j = -\frac{p_{j(v)}}{\frac{\partial p_{j(v)}}{\partial v_j}} = \frac{1}{1-p_j(v)}$$

Dessa forma, o mark-up sobre o custo marginal depende apenas do market share, e de modo crescente! A implicação é que em mercados especializados sem substitutos próximos, os mark-ups seriam pequenos, ao contrário do que seria esperado.

[5] *Trata-se apenas de uma normalização da função utilidade, que como sabemos, tem interpretação apenas ordinal.*

2.6 Contornando a Independência de Alternativas Irrelevantes

2.6.1 Probit multinomial

No caso geral, evita o problema de IIA por permitir correlações não triviais entre os não-observáveis. O custo deste procedimento é ter que computar a distribuição acumulada conjunta dos não-observáveis, avaliada no ponto coerente com as decisões dos agentes:

$$\Pr\left(d_{ij}=1|X_i\right)=\Pi_{j=1}^{J}\Phi\left[\left(\gamma_1-\gamma_j\right)'X_i,\left(\gamma_2-\gamma_j\right)'X_i,...,\left(\gamma_J-\gamma_j\right)'X_i;A_j'\Sigma_\varepsilon A_j\right]^{d_{ij}}$$

$$\begin{aligned}&\Phi\left[\left(\gamma_1-\gamma_j\right)'X_i,\left(\gamma_2-\gamma_j\right)'X_i,...,\left(\gamma_J-\gamma_j\right)'X_i;A_j'\Sigma_\varepsilon A_j\right]\\ =&\int_{-\infty}^{\left(\gamma_1-\gamma_j\right)'X_i}...\int_{-\infty}^{\left(\gamma_J-\gamma_j\right)'X_i}\phi\left[\left(\gamma_1-\gamma_j\right)'X_i,\left(\gamma_2-\gamma_j\right)'X_i,...,\left(\gamma_J-\gamma_j\right)'X_i;A_j\Sigma_\varepsilon A_j'\right]d\varepsilon_1,...,d\varepsilon_J\end{aligned}$$

onde A_j é uma matriz com 1's na diagonal, -1's na coluna j, e zeros nas demais coordenadas.

A principal dificuldade é computar a integral múltipla acima, e tipicamente não se consegue estimar um modelo destes por máxima verossimilhança num computador caseiro, se houver mais de 5 escolhas simultâneas (ou até menos, se a base de dados for grande. Grosso modo, o custo computacional cresce exponencialmente com o número de escolhas e linearmente com o número de observações). Alternativas à máxima verossimilhança incluem (i) utilizar métodos menos precisos de aproximação numérica da integral acima (comumente, utilizam-se quadraturas gaussianas) podem aumentar a capacidade ao custo de menor acurácia, (ii) métodos que involvem simulação como Máxima Verossimilhança Simulada ou Método dos Momentos Generalizado Simulado demandam em geral menos capacidade computacional,

mas perdem acurácia e eventualmente produzem estimadores inconsistentes, (iii) métodos Bayesianos baseados em simulação, que além de tudo evitam maximização. Recentemente o simulador de probabilidades GHK parece ter resolvido boa parte dos problemas, oferecendo acurácia a um baixo custo computacional, expandindo a dimensionalidade factível do problema para mais dimensões (mas não muitas mais).

2.6.2 Logit sequencial ou em níveis (nested logit)

O modelo de decisões sequenciais pode resolver o problema anterior mudando a estrutura decisória dos agentes. No exemplo acima, imagine que os agentes primeiro decidem se querem um sedan ou um esportivo, e para aqueles que escolherem o esportivo surge outro nó decisório onde têm que optar entre verde e vermelho. Desse modo, podemos inserir alguma correlação entre os não-observáveis sem fugir do setup tratável do logit. Em termos da probabilidade conjunta dos não-observáveis, a hipótese central do modelo é a de que:

$$F\left(\varepsilon_1,...,\varepsilon_J\right) = \exp - \left[e^{-\left(\frac{\varepsilon_1+\varepsilon_2+...+\varepsilon_{J_1}}{\sigma_1}\right)^{\sigma_1}} e^{-\left(\frac{\varepsilon_{J_1+1}+...+\varepsilon_{J_2}}{\sigma_2}\right)^{\sigma_2}} ... e^{-\left(\frac{\varepsilon_{J_{M-1}+1}+...+\varepsilon_{J_M}}{\sigma_M}\right)^{\sigma_M}} \right]$$

No primeiro momento, o agente decide um dos M subconjuntos de escolhas, ao passo que no segundo momento ele escolhe uma das escolhas dentro do grupo (o modelo generaliza para mais níveis). Uma forma alternativa de escrever as probabilidades do modelo a partir dessa estratégia é criar uma variável categórica $C = 1, ..., M$ denotando as opções no primeiro estágio, e escrever (por simplicidade, omiti o fato de que todas as probabilidades abaixo

estão condicionadas nos observáveis, X).

$$\Pr(d_j = 1) = \Pr(d_j = 1 | j \in c \subset C) \Pr(c)$$

Para utilizar esta estratégia, é necessário fazer algumas normalizações adicionais. Se considerarmos somente as alternativas disponíveis dentro de um determinado grupo c, temos:

$$\begin{aligned} v_{1c} &= \beta'_{1c} X \\ &\vdots \\ v_{Jc} &= \beta'_{Jc} X \end{aligned}$$

definindo $\gamma_{jc} = \beta_{jc} - \beta_{1c}$:

$$\begin{aligned} v_{1c} &= \beta'_{1c} X \\ &\vdots \\ v_{Jc} &= \left(\beta_{1c} + \gamma_{jc}\right)' X \end{aligned}$$

Neste caso, as duas componentes da probabilidade que formam $\Pr(d_j = 1)$ são:

$$\begin{aligned} \Pr(d_j = 1 | j \in c \subset C) &= \frac{\exp\left(\beta'_{jc} X/\sigma_c\right)}{\sum_{jc=1}^{J_c} \exp\left(\beta'_{jc} X/\sigma_c\right)} \\ &= \frac{e^{\beta'_{1c} X/\sigma_c}}{e^{\beta'_{1c} X/\sigma_c}} \frac{\exp\left(\gamma'_{jc} X/\sigma_c\right)}{\sum_{jc=1}^{J_c} \exp\left(\gamma'_{jc} X/\sigma_c\right)} \\ &= \frac{\exp\left(\gamma'_{jc} X/\sigma_c\right)}{\sum_{jc=1}^{J_c} \exp\left(\gamma'_{jc} X/\sigma_c\right)} \end{aligned}$$

onde $\gamma_{1c} = 0$, e essa passa a ser mais uma normalização necessária (ou em outras palavras, podemos identificar γ_{jc}, definido em termos de uma alternativa de referência dentro do grupo c, implicando em c normalizações ao invés de apenas uma).

$$\begin{aligned}
\Pr(c) &= \frac{e^{\sigma_c I_c + \alpha_c}}{\sum_{c=1}^{M} e^{\sigma_c I_c + \alpha_c}} \\
I_c &= \ln\left[\sum_{j \in c} \frac{\exp\left(\gamma'_{jc} X\right)}{\sigma_c}\right] \\
\alpha_c &= \sum_{j \in c} \gamma'_{jc} X
\end{aligned}$$

2.7 Modelos condicionais

Até o momento, supusemos que os benefícios líquidos associados a cada escolha, $V_{ij}(X_i, U_{ij})$ dependiam apenas de características individuais, isto é, agentes com características observáveis (X_i, U_{ij}) optavam entre uma de $j = 0, ..., J$ escolhas segundo o critério

$$\begin{aligned}
d_{ij} &= 1 \Leftrightarrow V_{ij}(X_i, U_{ij}) > \max_{k \neq j} V_{ik}(X_i, U_{ik}) \\
V_{ij}(X_i, U_{ij}) &= v_{ij}(X_i) + U_{ij} \\
v_{ij}(X_i) &= \beta'_j X_i
\end{aligned}$$

Em muitas situações, contudo, observamos não apenas características dos indivíduos mas também atributos das escolhas. Se nosso problema por exemplo for modelar a escolha de meio de transporte de um indivíduo que tem que se deslocar de casa ao trabalho, uma variável que pode ser fundamental na decisão é o tempo gasto por cada meio de transporte, e essa tipicamente é uma característica da escolha e não do agente. Similarmente, características

do agente que afetem os benefícios líquidos de apenas uma das escolhas envolvidas podem ser modeladas como sendo características da escolha, já que é alguma particularidade da escolha que faz com que aquela característica seja importante para ela e não para as demais opções (por exemplo, na escolha de carreira profissional suponha que temos as opções médico e engenheiro, e que uma das características do agente seja uma dummy indicando se o indivíduo é filho de médico. Podemos tanto dizer que essa é de fato uma característica do agente - ter nascido filho de médico - , quanto uma especificidade da carreira de medicina, que é a única para a qual ser filho de médico tem alguma importância).

A forma de modelar características da escolha num modelo condicional é:

$$V_{ij} = \delta' Z_{ij} + U_{ij}$$

Note que, neste caso, os coeficientes não variam entre alternativas, mas os regressores, sim. Desse modo, no exemplo da escolha de transporte, temos duas variáveis distintas para "duração da viagem de ônibus" e "duração da viagem de carro", mas apenas um coeficiente capturando a desutilidade marginal de demorar um segundo a mais para chegar no trabalho. No caso da escolha profissional, a variável "ser filho de médico na carreira de medicina" é uma dummy, ao mesmo tempo que essa variável não entra no cômputo do benefício líquido associado a engenharia.

A primeira diferença que pode ser notada entre variáveis inerentes à escolha e variáveis comuns às escolhas é que no primeiro caso o coeficiente δ pode ser diretamente estimado (a menos da escala) num problema de escolha discreta, ao passo que no caso de β podemos

apenas estimar $\gamma_j = \beta_j - \beta_0$, para alguma alternativa de referência 0.

Antes de prosseguir nossa análise, vale ressaltar que nada impede o analista de incluir ambos regressores específicos da escolha e regressores específicos do agente no modelo, tendo como caso geral:

$$V_{ij} = \beta_j' X_i + \delta' Z_{ij} + U_{ij}$$

A segunda característica importante (e de onde surge o termo "condicional") é que a forma de estimação é baseada em log-verossimilhança condicional. Condicional em que? Originalmente estes modelos foram propostos para lidar com bases de dados longitudinais ("em painel"), onde um mesmo agente é observado por diversas vezes fazendo mais de uma escolha. Nestes modelos, é comum supor que há ao menos um componente não-observável do agente que persiste ao longo do tempo, gerando correlação entre as decisões. Uma das formas de lidar com este componente persistente é tentar encontrar alguma variável observada que seja *estatística suficiente* para ele, isto é, uma variável tal que, uma vez que a função de verossimilhança esteja condicionada nela, o componente persistente não-observável deixa de afetar as decisões dos indivíduos. No caso mais comum, o do logit condicional, a estatística suficiente é a soma de 1's nas decisões dos agentes (isto é, o número total de vezes que o agente decidiu em favor da opção 1 em detrimento de 0). Alternativamente, poderia-se condicionar na média da variável dependente d, que é simplesmente a soma de 1's dividida por uma constante (o número de vezes que o indivíduo aparece na amostra).

O problema é que condicionando no número de decisões não-nulas do agente, todos os regressores que não variarem entre alternativas se cancelam na função de verossimilhança, de

modo que fica impossível estimar os coeficientes associados a eles. Para ver isso, vamos considerar dois casos: (i) uma escolha binária repetida duas vezes no tempo, e (ii) uma escolha entre três alternativas em um ponto do tempo (no caso de uma escolha binária observada uma única vez, é trivial ver que se condicionarmos no número de vezes em que $d = 1$, o modelo fica degenerado e prediz perfeitamente a variável d, sem precisar da ajuda de nenhum outro regressor. No caso de mais alternativas/ períodos a generalização é automática)

(i) suponha que observemos o agente duas vezes, e queremos estimar a probabilidade de que na primeira observação a escolha tenha sido 0 e na segunda 1, condicional no fato de que em ao menos uma das observações a escolha foi 1. Usando a regra de Bayes, temos:

$$\begin{aligned} & \Pr\left[d_{i1} = 0 \ \& \ d_{i2} = 1 | d_{i1} + d_{i2} = 1, X\right] \\ = & \frac{\Pr\left(d_{i1} = 0|X\right) \Pr\left(d_{i2} = 1|X\right)}{\Pr\left(d_{i1} = 0|X\right) \Pr\left(d_{i2} = 1|X\right) + \Pr\left(d_{i1} = 1|X\right) \Pr\left(d_{i2} = 0|X\right)} \end{aligned}$$

como no caso do logit $\Pr\left(d_{it} = 1|X\right) = \frac{e^{\gamma' X_{it}}}{1+e^{\gamma' X_{it}}} = \frac{e^{\gamma' X_{it} + \overline{\gamma}' \overline{X}_i}}{1+e^{\gamma' X_{it} + \overline{\gamma}' \overline{X}_i}}$ (onde as variáveis com a barra são aquelas que não variam no tempo):

$$\begin{aligned} & \Pr\left[d_{i1} = 0 \ \& \ d_{i2} = 1 | d_{i1} + d_{i2} = 1, X\right] \\ = & \frac{\frac{1}{1+e^{\gamma' X_{i1} + \overline{\gamma}' \overline{X}_i}} \frac{e^{\gamma' X_{i2} + \overline{\gamma}' \overline{X}_i}}{1+e^{\gamma' X_{i2} + \overline{\gamma}' \overline{X}_i}}}{\frac{1}{1+e^{\gamma' X_{i1} + \overline{\gamma}' \overline{X}_i}} \frac{e^{\gamma' X_{i2} + \overline{\gamma}' \overline{X}_i}}{1+e^{\gamma' X_{i2} + \overline{\gamma}' \overline{X}_i}} + \frac{1}{1+e^{\gamma' X_{i2} + \overline{\gamma}' \overline{X}_i}} \frac{e^{\gamma' X_{i1} + \overline{\gamma}' \overline{X}_i}}{1+e^{\gamma' X_{i1} + \overline{\gamma}' \overline{X}_i}}} \\ = & \frac{e^{\gamma' X_{i2} + \overline{\gamma}' \overline{X}_i}}{e^{\gamma' X_{i2} + \overline{\gamma}' \overline{X}_i} + e^{\gamma' X_{i1} + \overline{\gamma}' \overline{X}_i}} \\ = & \frac{e^{\overline{\gamma}' \overline{X}_i}}{e^{\overline{\gamma}' \overline{X}_i}} \frac{e^{\gamma' X_{i2}}}{e^{\gamma' X_{i2}} + e^{\gamma' X_{i1}}} = \frac{e^{\gamma' X_{i2}}}{e^{\gamma' X_{i2}} + e^{\gamma' X_{i1}}} \end{aligned}$$

(ii) No caso de múltipla escolha, para seguir o mesmo raciocínio devemos definir as variáveis

$$\begin{aligned}
\widetilde{d}_1 &= 1\left[v_1 + U_1 > v_0 + U_0\right] \\
\widetilde{d}_2 &= 1\left[v_2 + U_2 > v_0 + U_0\right] \\
\widetilde{d}_3 &= 1\left[v_2 + U_2 > v_1 + U_1\right]
\end{aligned}$$

e condicionamos portanto em $\sum_j \widetilde{d}_j$. Posto dessa forma,

$$\begin{aligned}
&\Pr\left(d_{i2} = 1 | X, \sum_j \widetilde{d}_j = 2\right) \\
= \quad & \frac{\Pr\left(\widetilde{d}_2 = 1|X\right)\Pr\left(\widetilde{d}_3 = 1|X\right)}{\Pr\left(\widetilde{d}_1 = 0|X\right)\Pr\left(\widetilde{d}_2 = 1|X\right)\Pr\left(\widetilde{d}_3 = 1|X\right)} \\
& + \Pr\left(\widetilde{d}_1 = 1|X\right)\Pr\left(\widetilde{d}_2 = 1|X\right)\Pr\left(\widetilde{d}_3 = 0|X\right) \\
& \Pr\left(\widetilde{d}_1 = 1|X\right)\Pr\left(\widetilde{d}_2 = 0|X\right)\Pr\left(\widetilde{d}_3 = 1|X\right)
\end{aligned}$$

e algebra similar ao exemplo anterior mostra que somente os termos associados a variáveis que não são constantes entre escolhas permanecerão na verossimilhança após o condicionamento.

É importante notar, contudo, que no caso de múltiplas escolhas podemos "driblar" a "multicolinearidade" decorrente do cancelamento no numerador e denominador dos termos invariantes, através de um truque simples. Primeiramente, definimos novas variáveis como sendo interações dos regressores invariantes com as dummies referentes à escolha. Assim, se

as escolhas são "Viajar para Veneza", "Viajar para a Disney" ou "Viajar para Nova Iguaçú", e o regressor invariante é idade, construímos novas variáveis como "idade, caso viaje para Veneza", "idade, caso viage para Nova Iguaçu", etc.

Como previsto, um modelo estático condicional usando estes truques deveria fornecer resultado semelhante ao que teríamos caso estimássemos diretamente um logit multinomial, e de fato é isso que ocorre. Essencialmente, os modelos condicionais desempenham papel importante quando usamos explicitamente a longitudinalidade dos dados, permitindo como foi dito controlar para efeitos não-observáveis persistentes.

2.8 Probit e logit ordenados

Há um caso particular de escolhas múltiplas em que existe um critério óbvio para "ranquear" as opções disponíveis, de modo a simplificar a estimação para um conjunto particular de hipóteses adicionais.

Considere o caso em que um indivíduo tem que escolher entre 3 loterias, onde todas oferecem a mesma recompensa média, mas podem ser ordenadas segundo o risco envolvido (a primeira com variância pequena, a segunda com variância média e a última com variância grande). Suponha ainda que o componente individual não-observado seja o grau de aversão a risco dos indivíduos, e que o custo dos bilhetes destas loterias seja decrescente no risco. Formalmente, seja α o prêmio médio pago por cada loteria, c_j o custo (não-observado e constante) de participar da loteria, e U_i o grau de aversão a risco do agente. O o problema pode ser representado da seguinte forma:

$$
\begin{aligned}
v_1 &= \alpha - c_1 - \sigma_1 U_i \\
v_2 &= \alpha - c_2 - \sigma_2 U_i \\
v_3 &= \alpha - c_3 - \sigma_3 U_i
\end{aligned}
$$

$$
\begin{aligned}
d_{i1} &= 1 \Leftrightarrow U_i \leq \frac{c_2 - c_1}{\sigma_1 - \sigma_2} \\
d_{i1} &= 2 \Leftrightarrow \frac{c_2 - c_1}{\sigma_1 - \sigma_2} < U_i \leq \frac{c_3 - c_2}{\sigma_2 - \sigma_3} \\
d_{i1} &= 3 \Leftrightarrow \frac{c_3 - c_2}{\sigma_2 - \sigma_3} < U_i
\end{aligned}
$$

Se em nosso modelo a distribuição de U_i é F_U, temos que:

$$
\begin{aligned}
\Pr\left(d_{i1} = 1\right) &= F\left(\frac{c_2 - c_1}{\sigma_1 - \sigma_2}\right) \\
\Pr\left(d_{i1} = 1\right) &= F\left(\frac{c_3 - c_2}{\sigma_2 - \sigma_3}\right) - F\left(\frac{c_2 - c_1}{\sigma_1 - \sigma_2}\right) \\
\Pr\left(d_{i1} = 1\right) &= 1 - F\left(\frac{c_2 - c_1}{\sigma_1 - \sigma_2}\right) - F\left(\frac{c_3 - c_2}{\sigma_2 - \sigma_3}\right) \\
\pounds\left(\beta\right) &= \prod_{i=1}^{N} \begin{bmatrix} F\left(\frac{c_2 - c_1}{\sigma_1 - \sigma_2}\right)^{d_{i1}} \\ \left[F\left(\frac{c_3 - c_2}{\sigma_2 - \sigma_3}\right) - F\left(\frac{c_2 - c_1}{\sigma_1 - \sigma_2}\right)\right]^{d_{i2}} \\ \left[1 - F\left(\frac{c_2 - c_1}{\sigma_1 - \sigma_2}\right) - F\left(\frac{c_3 - c_2}{\sigma_2 - \sigma_3}\right)\right]^{1 - d_{i1} - d_{i2}} \end{bmatrix}
\end{aligned}
$$

e podemos estimar as combinações de parâmetros $\frac{c_2 - c_1}{\sigma_1 - \sigma_2}$, $\frac{c_2 - c_1}{\sigma_1 - \sigma_2}$, $\frac{c_3 - c_2}{\sigma_2 - \sigma_3}$, ou c_1, c_2, c_3 após normalizações das variâncias.

De um modo geral, podemos incluir heterogeneidade observada também no modelo (juntamente com a heterogeneidade não observada expressa em U_i), desde que este possa ser rescrito da seguinte forma:

$$\begin{aligned} v_{ij} &= \alpha_j + \beta' X + U_i \\ &= \alpha_j + v_{ij}^* \\ d_{i0} &= 1 \Leftrightarrow v_{ij}^* \leq \alpha_0 \\ d_{i1} &= 1 \Leftrightarrow \alpha_0 < v_{ij}^* \leq \alpha_1 \\ &\vdots \\ d_{iJ} &= 1 \Leftrightarrow \alpha_J < v_{ij}^* \end{aligned}$$

A função de verossimilhança nesse caso é:

$$\pounds(\beta) = \prod_{i=1}^{N} \left[\begin{array}{c} F(\alpha_0 - \beta' X)^{d_{i0}} * \\ \prod_{j=1}^{J-1} \left[F(\alpha_j - \beta' X) - F(\alpha_{j-1} - \beta' X) \right]^{d_{ij}} * \\ \left[1 - \sum_{j=0}^{J-1} F(\alpha_0 - \beta' X) \right]^{1 - \sum_{j=0}^{J-1} d_{ij}} \end{array} \right]$$

2.9 Outros comentários

2.9.1 Sobre a normalização de coeficientes

Vimos que de um modo geral, os objetos que podem ser identificados numa estimação de um modelo com variáveis dependentes discretas são as diferenças de utilidades normalizadas, $H_{jk}(X_i) = [v_j(X_i) - v_k(X_i)] / \sigma_j$. Isso porque (i) no processo decisório dos agentes apenas a

diferença entre recompensas importa para a tomada de decisões, e tudo o que observamos são as decisões (na seção seguinte veremos o que acontece quando também podemos observar algo sobre as recompensas), e (ii) o sinal da diferença de recompensas não se altera se multiplicarmos esta diferença por uma constante positiva, σ_j.

Há contudo um caso interessante que merece ser analisado com mais detalhe. Suponha que para cada escolha j, a utilidade v_j represente um benefício líquido associado à escolha j, e que este seja do tipo $v_j = \pi_j - c_j$, com π capturando o benefício líquido (que pode ter uma parte medida em reais, como por exemplo o lucro associado a esta escolha, e outra parte não monetária - mas passível de ser transformada em reais - como por exemplo a satisfação ou status de pertencer ao grupo dos que fizeram a escolha j), e c medindo os custos de escolher j. Ora, se c_j (ou $c_j - c_k$) for observável, então podemos de antemão fixar o coeficiente associado a esta variável em 1, e isto implicitamente determina uma escala de normalização, de modo que neste caso podemos estimar σ_j conjuntamente com os demais parâmetros do modelo. Em outras palavras, através da variação de uma determinada covariada X_{ik}, podemos em geral estimar $\widetilde{\gamma}_{jk} = \left(\frac{\gamma_{jk}}{\sigma_j}\right); \gamma_j = \beta_j - \beta_0$, para alguma alternativa de referência 0. Se em nossa regressão soubermos de antemão que $\gamma_{jk} = 1$ para a variável $X_k = c_j - c_0$, então quando analisarmos o coeficiente estimado $\widetilde{\gamma}_{j(c_j - c_0)} = \frac{1}{\sigma_j}$, teremos determinado o valor de σ_j, e a partir disso podemos inferir todos os demais valores de γ_{jk} fazendo $\gamma_{jk} = \frac{\widetilde{\gamma}_{jk}}{\widetilde{\gamma}_{j\left(c_j - c_0\right)}}$. Eventualmente será necessário restringir o coeficiente $\widetilde{\gamma}_{j(c_j - c_0)}$ para ser positivo em nossa regressão, já que a variância não pode ser negativa ou nula neste caso (alternativamente podemos impor já na função de verossimilhança que $\gamma_{j(c_j - c_0)} = 1$, e maximizar também em

σ).

2.9.2 Previsões

Grosso modo, há duas formas de utilizar as estimativas de um modelo de escolha discreta para fazer previsões. Considere o resultado de um modelo que visa estimar $\Pr(d_i = 1|X_i = x)$, com decisão binária e um único regressor X_i (a mesma lógica vale para múltiplas escolhas e regressores). Chame a probabilidade predita de $\widehat{P}_i(x)$. O modelo prediz portanto que o número de indivíduos que escolhem $d = 1$, é:

$$\widehat{N}_1 = \sum_{i=1}^{N} w_i \widehat{P}_i \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)$$

onde w_i é o peso amostral da observação i. A partir deste fato, há dois tipos de exercícios de previsão que podem ser feitos. No primeiro, gostaríamos de saber qual o número de pessoas que escolheriam $d = 1$ caso a população se mantivesse constante (isto é, caso os pesos amostrais não variassem), mas a distribuição das características que determinam a escolha dos agentes (neste caso X), sim. Neste caso, cujo exemplo visto em sala foi aumentar em 1 ano a escolaridade dos homens, e a variável dependente sendo a decisão de participar da força de trabalho, devemos calcular:

$$\widehat{N}_1^{(0)} = \sum_{i=1}^{N} w_i \widehat{P}_i \left(\widehat{\beta}_0 + \widehat{\beta}_1 \widetilde{x}_i\right)$$

onde $\widetilde{x}$ representa a nova variável aleatória. Note que a previsão pode ter como objetivo tanto avaliar o efeito de uma mudança natural em x (por exemplo, envelhecimento, mudanças na estrutura da pirâmide demográfica, eliminação gradual da indexação após estabilização de

preços), quanto o efeito causado pela introdução de uma política (pública, empresarial, familiar, etc) sobre x. Neste último caso, há que se ter cuidado com (i) a possibilidade de x ser determinado endogenamente ao modelo, em decisões que levam em conta x e d simultaneamente, e (ii) a possibilidade de que o impacto da política sobre x não seja plenamente conhecido (neste caso pode-se trabalhar com cenários).

A segunda forma de previsão se dá sobre mudanças nos pesos amostrais, que em geral são baseados completamente por algum outro conjunto de variáveis, z, de nossa amostra. Em uma pesquisa domiciliar, tipicamente sexo, região, e eventualmente escolaridade e raça podem estar entre os componentes que definem os estratos amostrais, de modo que $w_i = w(z_i)$. Assim como fizemos anteriormente, calculando nossa previsão baseado em transformações na distribuição de x mantendo a distribuição de z constante, poderíamos fazer o oposto, alterando a distribuição de z e mantendo a de x constante.

Observação importante: em geral, x e z não precisam representar o mesmo conjunto de variáveis, mas nada impede que haja componentes de x que também pertençam a z e vice-versa. Um exercício que tenha como ingrediente a mudança na distribuição de uma variável que pertença aos dois conjuntos deve acomodar transformações tanto nos pesos quanto nas probabilidades preditas.

Por último, há um parâmetro diferente dos demais numa regressão: o intercepto, que mede basicamente a média dos efeitos não-observados sobre a variável dependente (e graças a isso podemos normalizar a média do resíduo para zero sem perda de generalidade). Ao fazer previsões, contudo, podemos querer não somente ajustar as distribuições dos observáveis, mas

também a média dos não-observáveis. Se após a previsão tivermos acesso a dados da região de previsão, um exercício interessante pode ser estimar qual a parcela da mudança de N_1 para N_1^* que pode de fato ser atribuída a mudanças em x. Uma forma de recalibrar o intercepto de modo a capturar ajustes do intercepto simultâneos a ajustes nos observáveis é usar a diferença entre a variação prevista pelo modelo, $\widehat{N}_1^{(0)}$, e a variação observada na realidade, N_1^*, para ajustar o modelo como se o erro de previsão pudesse ser então integralmente atribuído a mudanças na média dos não-observáveis. Um algorítmo simples que faz este trabalho para nós é:

(i) defina um critério de convergência, ϵ

(ii) sugira um valor alternativo para o intercepto, $\widehat{\beta}_0^{(1)} = \widehat{\beta}_0 + \ln\left(N_1^*/\widehat{N}_1^{(0)}\right)$

(iii) estime o número de indivíduos que escolheria $d = 1$ caso o modelo fosse $\widehat{P}_i\left(\widehat{\beta}_0^{(1)} + \widehat{\beta}_1\widetilde{x}_i\right)$.

$$\widehat{N}_1^{(1)} = \sum_{i=1}^{N} w_i \widehat{P}_i\left(\widehat{\beta}_0^{(1)} + \widehat{\beta}_1\widetilde{x}_i\right)$$

(iv) compare $\widehat{N}_1^{(1)}$ e N_1^*. Se a diferença for menor que ϵ, pare o algorítmo e redefina o intercepto como $\widehat{\beta}_0^{(1)}$. Caso contrário, sugira um novo valor $\widehat{\beta}_0^{(2)} = \widehat{\beta}_0^{(1)} + \ln\left(N_1^*/\widehat{N}_1^{(1)}\right)$, e repita as etapas (ii) e (iii) até que $\widehat{N}_1^{(m)} - N_1^* < \epsilon$.

Finalmente, utilize o modelo ajustado $\Pr\left(d_i = 1 | X_i = x\right) = F\left(\widehat{\beta}_0^{(m)} + \widehat{\beta}_1 x\right)$ para estimar a contribuição efetiva de mudanças em x para as mudanças em N_1, fazendo

$$\sum_{i=1}^{N} w_i\left[\widehat{P}_i\left(\widehat{\beta}_0^{(m)} + \widehat{\beta}_1\widetilde{x}_i\right) - \left(\widehat{\beta}_0^{(m)} + \widehat{\beta}_1 x_i\right)\right]$$

3 Modelo de Roy de vantagens comparativas

3.1 Motivação

Considere uma economia com duas ocupações, 0 e 1, e com dois tipos de habilidade, S_0 e S_1, distribuídas na população segundo $f(s_0, s_1)$ no suporte dos reais positivos. Os preço associados a cada tipo de habilidade são π_0, π_1, e um indivíduo com $(S_0, S_1) = (s_0, s_1)$ escolhe a ocupação que maximiza seus rendimentos:

$$d = 1 \Leftrightarrow \pi_1 s_1 > \pi_0 s_0$$

A proporção de indivíduos que escolhe o setor 1 pode ser escrita como:

$$P_1 = \int_0^{\infty} \int_0^{\pi_1 s_1/\pi_0} f(s_0, s_1)\, ds_0 ds_1$$

Enquanto na população em geral a distribuição de habilidades do tipo 1 seria

$$f_1(s_1) = \int_0^{\infty} f(s_0, s_1)\, ds_0$$

, em nossa amostra poderíamos observar apenas

$$f_1^*(s_1|d = 1) = \frac{1}{P1} \int_0^{\pi_1 s_1/\pi_0} f(s_0, s_1)\, ds_0$$

Da mesma forma, para salários definidos como $w_j = \pi_j s_j$, teríamos a densidade marginal como:

$$g_1(w_1) = f_1\left(\frac{w_1}{\pi_1}\right)$$

e a condicional (observada):

$$g_1^*(w_1|d=1) = \frac{1}{\pi_1 P_1}\int_0^{w_1/\pi_0} f\left(s_0, \frac{w_1}{\pi_1}\right) ds_0$$

Ao problema relacionado ao fato de não observarmos $g_j(w_j)$ diretamente nos dados dá-se o nome de problema de seletividade (pois não observamos a distribuição incondicional de w, mas sim a distribuição condicional no fato de que w_j só é observado para os indivíduos que se auto-selecionaram no setor j). De forma análoga, estimativas de momentos populacionais $M_{kj}(w_j)$ baseados na hipótese de que a distribuição $g_j(w_j)$ é diretamente observada, são potencialmente viesadas se a distribuição realmente observada for $g_1^*(w_1|d=1)$. A esse viés dá-se o nome de viés de seleção.

A densidade observada de salários é dada por:

$$g(w) = P_1 g_1^*(w_1|d=1) + (1-P_1) g_0^*(w_0|d=0)$$

Na versão simplificada, supõe-se que $f(s_0, s_1)$ é log-normal, isto é:

$$\begin{pmatrix} \ln s_0 \\ \ln s_1 \end{pmatrix} \sim N\left[\underbrace{\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{pmatrix}}_{\Sigma}\right]$$

Como compatibilizar essa notação com o que tínhamos visto antes? Ora, neste caso $\mu_j = X'\beta_j$ (sem intercepto) e $s_j = \mu_j + U_j$, onde $U \sim N(0, \Sigma)$. O log do salário pode ser escrito como:

$$\ln w_j = \ln \pi_j + X'\beta_j + U_j$$

O problema neste caso é que o termo não observado não apenas não tem média zero na amostra (pois sua média é zero somente na distribuição incondicional, mas o que observamos já é condicional na escolha realizada), como tampouco é independente (ou independente-em-média) de X, pois mesmo que incondicionalmente o seja, quando condicionamos no evento $d = 1$ (ou $d = 0$), ele deixa de ser (pois X está entre os determinantes de d).

Como proceder neste caso?

Defina:

$$\begin{aligned}
\varepsilon &= U_1 - U_0 \\
\sigma^* &= \sqrt{var(\varepsilon)} = \sqrt{\sigma_{00} - 2\sigma_{10} + \sigma_{11}} \\
c &= \ln\left(\frac{\pi_1}{\pi_0}\right) + X'(\beta_1 - \beta_0) \\
c^* &= \frac{c}{\sigma^*} \\
P_1 &= \Pr(\ln w_1 > \ln w_0) = 1 - \Phi(-c^*) = \Phi(c^*)
\end{aligned}$$

Temos então que:

$$E\left[\ln w_1 | \ln w_1 > \ln w_2, X\right] = \ln \pi_1 + X'\beta_1 + E\left[U_1 | \ln w_1 > \ln w_2, X\right]$$

Como resultado da normalidade, temos ainda que (U_1, ε) são distribuídos segundo uma normal bivariada, e que:

$$\begin{aligned}
U_1 &= a_1\varepsilon + V_1 \\
a_1 &= \frac{cov\left(U_1, \varepsilon\right)}{var\left(\varepsilon\right)} = \frac{\sigma_{11} - \sigma_{01}}{\left(\sigma^*\right)^2} \\
V_1 &\sim N\left(0, \sigma_{11}\left(1 - \rho_1^2\right)\right) \\
\rho_1 &= \frac{cov\left(U_1, \varepsilon\right)}{\sqrt{var\left(\varepsilon\right) var\left(U_1\right)}} = \frac{\sigma_{11} - \sigma_{01}}{\sigma^*\sqrt{\sigma_{11}}}
\end{aligned}$$

Mais do que isso, essa decomposição garante que V_1 é independente de ε (e portanto, independente de d, já que $d = 1\left[\varepsilon > -c\right]$). Nossa equação de regressão fica assim:

$$\begin{aligned}
E\left[\ln w_1 | \ln w_1 > \ln w_2, X\right] &= \ln \pi_1 + X'\beta_1 + a_1 E\left[\varepsilon | \varepsilon > -c\right] \\
&= \ln \pi_1 + X'\beta_1 + \sigma^* a_1 E\left[z | z > -c^*\right] \\
&= \ln \pi_1 + X'\beta_1 + \frac{\sigma_{11} - \sigma_{01}}{\sigma^*} E\left[z | z > -c^*\right]
\end{aligned}$$

onde z denota uma variável aleatória normal padrão. Usando estratégia semelhante, podemos mostrar também que:

$$var\left[\ln w_1 | \ln w_1 > \ln w_0\right] = \sigma_{11}\left[\rho_1^2 var\left(z | z > -c^*\right) + \left(1 - \rho_1^2\right)\right]$$

Alguns fatos sobre distribuições normais:

- $E\left[z|z>c\right]=\frac{(2\pi)^{-1/2}e^{-\frac{c^2}{2}}}{\Phi(-c)}=\lambda\left(c\right)=\frac{\phi(-c)}{\Phi(-c)}$
- $var\left(z|z>c\right)=1+c\lambda\left(c\right)-\left[\lambda\left(c\right)\right]^2$
- $\lim_{c\to\infty}\lambda\left(c\right)=\infty;\ \lim_{c\to-\infty}\lambda\left(c\right)=0$
- $0<\frac{\partial\lambda(c)}{\partial c}=\lambda'\left(c\right)=\lambda\left(c\right)\left[\lambda\left(c\right)-c\right]<1$
- $\lim_{c\to\infty}\frac{\partial\lambda(c)}{\partial c}=1;\ \lim_{c\to-\infty}\frac{\partial\lambda(c)}{\partial c}=0$
- $0<\frac{\partial^2\lambda(c)}{\partial c^2}$
- $0>\frac{\partial var(z|z>c)}{\partial c}$
- $\lim_{c\to\infty}var\left(z|z>c\right)=0;\ \lim_{c\to-\infty}var\left(z|z>c\right)=1$
- Média de z $\geq$ Moda de z (condicional em z > c)

$$E\left[\ln w_1|\ln w_1>\ln w_2,X\right]=\ln\pi_1+X'\beta_1+\frac{\sigma_{11}-\sigma_{01}}{\sigma^*}\lambda\left(-c^*\right)$$

3.2 Implicações do modelo

1. Não pode ocorrer de ambas as variâncias serem menores que a covariância.

$$0\leq\left(\frac{\sigma_{01}}{\sigma_{11}}\right)\left(\frac{\sigma_{01}}{\sigma_{00}}\right)\leq1$$

2. Considere um aumento em π_1. O efeito deste aumento sobre a média condicional das habilidades nos dois setores é indeterminado:

$$\begin{aligned}\frac{\partial E\left[\ln s_1 | \ln w_1 > \ln w_2\right]}{\partial \ln \pi_1} &= -\frac{\sigma_{11}-\sigma_{01}}{\left(\sigma^*\right)^2}\lambda'\left(-c^*\right) \\ &= -\frac{\sigma_{11}-\sigma_{01}}{\left(\sigma^*\right)^2}\left[\lambda\left(-c^*\right)\left(\lambda\left(-c^*\right)+c^*\right)\right]\end{aligned}$$

É possível que haja até mesmo diminuição no salário pago em ambos os setores, se o efeito-composição superar o efeito-preço. No efeito-composição, parte das pessoas se deslocam do setor 0 para o setor 1. Se parte significativa dos que se deslocarem tiver elevada habilidade do tipo 0 e habilidade não tão grande do tipo 1, o salário médio se reduzirá nos dois setores (mas não a média salarial na população como um todo). Este será o caso se:

$$\frac{\sigma_{11}-\sigma_{01}}{\sigma^*}\lambda'\left(-c^*\right) > 1$$

o que é implicado por $\sigma_{00} < \sigma_{01}$.Além disso, se essa condição valer teremos que a auto-seleção faz com que a média de habilidade no setor zero seja menor do que a média incondicional, μ_0. Esse caso é denominado "não-usual", e não pode ocorrer em ambos os setores simultaneamente (pela implicação 1)

3. Uma economia de Roy possui menor variância de salários que uma economia com a mesma quantidade de pessoas em cada setor, porém distribuída aleatoriamente (segue de análise da equação da variância condicional e dos fatos sobre distribuições normais). Além disso, a auto-seleção reduz tanto a dispersão (medida pela variância do log do salário) agregada quanto a dispersão em cada setor da economia, se comparado à alocação aleatória de trabalhadores.

4. A cauda da distribuição de salários na economia de Roy é menos pesada que a cauda de uma densidade de Pareto.

5. Auto-seleção eleva a média de habilidade no setor j se $\sigma_{00} > \sigma_{01}$

6. Um aumento de π_1 faz com que a variância do log do salário no setor 1 aumente e no setor 2 diminua.

7. No caso "usual", a distribuição do log dos salários será assimétrica para a direita (similar a log-normal). No caso "não-usual", a assimetria será para a esquerda.

3.3 Identificação

No modelo de Roy, todos os parâmetros β_j, Σ podem ser identificados se os dados disponíveis contiverem observações sobre (i) salários, (ii) setor em que os agentes trabalham, e (iii) covariadas X.

Se não observarmos o setor em que as pessoas trabalham, podemos identificar os parâmetros mas não seus subscritos (isto é, não conseguimos saber se um determinado vetor $\widetilde{\beta}$ corresponde a β_1 ou β_0, o mesmo ocorrendo com as variâncias).

Se observarmos apenas os salários em um dos setores e as decisões dos agentes de participar ou não daquele setor (por exemplo, $w_1|w_1 > w_0$ e $\Pr(w_1 > w_0)$, então podemos identificar $\beta_1, \sigma_{11}, \sigma_{01}$, e β_0/σ_0 (ou outras combinações de β_0 e σ_0).

Com dados em painel, pode ser possível identificar mais parâmetros, dependendo das hipóteses feitas sobre o comportamento decisório dinâmico dos agentes.

3.4 Estimação

No caso do modelo de Roy, há basicamente duas formas de estimação comumente usadas. A primeira repete o que foi visto anteriormente no curso, e consiste em maximizar a função de log-verossimilhança, que neste caso é:

$$\begin{aligned}
\pounds(\theta) &= \prod_{i=1}^{N} h(d_i, w_{i1}, w_{i2}) \\
&= \prod_{i=1}^{N} \left\{ \begin{array}{c} [g_1(w_{i1}|X_i, d_i=1)\Pr(d_i=1|X_i)]^{d_i} \\ [g_0(w_{i0}|X_i, d_i=0)\Pr(d_i=0|X_i)]^{1-d_i} \end{array} \right\} \\
&= \prod_{i=1}^{N} \left\{ \begin{array}{c} \left[\frac{\phi(X_i'\beta_1)}{\Phi(-X_i'(\beta_1-\beta_0))}\Phi(-X_i'(\beta_1-\beta_0))\right]^{d_i} \\ \left[\frac{\phi(X_i'\beta_0)}{1-\Phi(-X_i'(\beta_1-\beta_0))}[1-\Phi(-X_i'(\beta_1-\beta_0))]\right]^{1-d_i} \end{array} \right\} \\
&= \prod_{i=1}^{N} \phi(X_i'\beta_1)^{d_i}\phi(X_i'\beta_0)^{1-d_i}
\end{aligned}$$

Um caso interessante ocorre quando observamos w_1 se $d=1$, mas não observamos w_0 quando $d=0$. Neste caso, a verossimilhança fica:

$$= \prod_{i=1}^{N} \phi(X_i'\beta_1)^{d_i}[1-\Phi(-X_i'(\beta_1-\beta_0))]^{1-d_i}$$

A forma alternativa é inspirada no método dos momentos e consiste em dois estágios. No primeiro estágio, estima-se um probit com a probabilidade de que o indivíduo escolha a alternativa 1 como alvo, isto é, estima-se

$$\Phi\left(-c_i^*\right) = \Phi\left[-\frac{1}{\sigma^*}\left(\ln\left(\frac{\pi_1}{\pi_0}\right) + X_i'\left(\beta_1 - \beta_0\right)\right)\right].$$

Note que a partir desta estimativa podemos inferir também $\phi\left(-c_i^*\right)$, que é simplesmente a derivada da primeira expressão com respeito a c^*, e também $\lambda\left(-c_i^*\right)$.

No segundo estágio, incluímos $Z_i = \lambda\left(-c_i^*\right)$ como regressor em nossa equação de salários, sabendo que

$$E\left[\ln w_1 | \ln w_1 > \ln w_2, X\right] = \ln \pi_1 + X'\beta_1 + \frac{\sigma_{11} - \sigma_{01}}{\sigma^*} Z_i$$

Ora, este é um momento amostral predito pelo modelo, e que pode ser diretamente utilizado para estimarmos os coeficientes $\frac{\sigma_{11}-\sigma_{01}}{\sigma^*}, \beta_1, \ln \pi_1$, utilizando, por exemplo, OLS. Este método em dois estágios é computacionalmente muito "barato", e permitiu a estimação de modelos de seleção mesmo nos anos 70, quando o estágio tecnológico dos computadores era relativamente atrasado se comparado aos dias atuais. Pela simplicidade, tais modelos se tornaram populares, e ganharam o nome de modelo de correção de Heckman, em homenagem ao seu propositor.

Há contudo uma importante dificuldade com a estimação do modelo de Roy em 2 estágios: no caso mais comumente estimado, observamos apenas o salário em um dos setores (em geral o salário dos que trabalham), e as escolhas dos agentes. Nesse caso, a variância no setor não observado deve ser normalizada para identificação, e nada garante que o coeficiente de correlação implicado pela estimação separada de σ_{11}, σ_{01} com σ_{00} normalizado para 1 fique no intervalo [-1,1]. No modelo estimado por máxima verossimilhança, essa restrição pode ser

imposta diretamente na estimação, e esse problema não ocorre.

3.5 Modelo de Roy generalizado

No caso generalizado, permite-se que as escolhas sejam feitas com base em critérios mais amplos que simplesmente a escolha do maior salário. Pode-se por exemplo imaginar que cada escolha envolva um custo $C_{ij} = X_i^c \gamma_j$, capturando o desprazer de exercer uma determinada profissão ou a distância do trabalho, etc. O modelo generalizado propõe que:

$$
\begin{aligned}
d_i &= 1 \Leftrightarrow I_i > 0 \\
I_i &= Z_i'\gamma + U_i^I \\
\begin{pmatrix} U_{i0} \\ U_{i1} \\ U_i^I \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{00} & \sigma_{01} & \sigma_{0I} \\ \sigma_{01} & \sigma_{11} & \sigma_{1I} \\ \sigma_{0I} & \sigma_{1I} & \sigma_{II} \end{pmatrix} \right] \\
w_j &= X_i'\beta_j + U_{ij}
\end{aligned}
$$

O raciocínio seguido para propor e resolver o modelo de Roy serve plenamente para o caso do modelo generalizado de Roy, exceto que neste é necessário a normalização de σ_{II} (em geral para 1), a menos que particularidades do modelo permitam identificar este parâmetro. Note que as variáveis presentes em Z e X não precisam coincidir, ainda que seja comum haver uma intersecção entre estes vetores. Assim como antes, modelo permite isolar a parte de U_{ij} correlacionada com X_i, via projeção de U_{ij} em U_{iI}, e usar os momentos condicionais derivados do modelo para fazer a estimação (ou alternativamente estimar tudo por máxima verossimilhança).

3.6 Tipos de amostra

Em nosso curso, focamos a análise na comparação entre dois tipos de amostra: (i) amostragem aleatória (pura e estratificada), e (ii) amostragem baseada em escolha. Suponha que na população estejamos interessados num conjunto de variáveis aleatórias com distribuição $f(X)$

Numa amostra aleatória pura, cada membro de uma dada população tem a mesma probabilidade de ser amostrado, e os sorteios que definem quem será amostrado são independentes entre si. Na amostra estratificada, divide-se a população em estratos e sorteia-se aleatoriamente indivíduos dentro de cada estrato. Neste caso, indivíduos de um mesmo estrato têm a mesma probabilidade de serem amostrados, mas pode haver estratos subamostrados ou super amostrados. Uma variante das amostras aleatórias são as amostras pseudo-aleatórias. No caso da PNAD por exemplo, a população é dividida em estratos e, dentro de cada estrato (composto por um conjunto de aproximadamente 100 domicílios chamado de setor censitário), sorteia-se apenas o primeiro domicílio a ser entrevistado e depois segue-se a regra de pular 7 domicílios e entrevistar o oitavo (numa rua, prédio, etc.). Se acreditarmos que a regra de entrevistar os múltiplos de 8 é independente das variáveis investigadas na pesquisa, a amostra terá propriedades semelhantes às de uma amostra estratificada. No caso da amostra aleatória simples, cada indivíduo representa uma realização de $X = x$, obtida com probabilidade $f(x)$. Na amostra estratificada, sabemos de antemão qua a probabilidade de que um indivíduo pertença a um estrato é $\Pr(E)$, e recuperamos f via $f_E(x|E)\Pr(E) = f(x)$.

Um segundo tipo de amostragem tem como base um subgrupo da população que possui uma dada característica definida a partir do comportamento dos participantes. Assim, uma

pesquisa eleitoral feita por telefone restringe a amostra aos donos de telefone, e se essa característica não for independente das informações que se quer coletar, as propriedades da amostra não se assemelharão às de uma amostra aleatória, e correções deverão ser feitas para que seja possível inferir algo estatisticamente deste grupo. Vimos que mesmo numa amostra aleatória nosso exercício pode estar restrito a um subgrupo amostrado de forma não aleatória, tal como o de pessoas ocupadas quando se quer fazer inferência sobre determinantes salariais. A abordagem de funções de controle fornece uma possível forma de correção. Tipicamente, indivíduos nesta amostra foram sorteados a partir de $f_D(x|D(x)=d)$, e possibilidades de correção dependem de quão razoáveis são as hipóteses sobre a relação $d(x)$.

Há ainda outros dois tipos de amostra frequentemente encontrados: (iii) amostras truncadas e (iv) amostras censuradas.

No caso das amostras truncadas, observamos indivíduos sorteados a partir de uma amostra truncada do tipo $f_T(a<X<b)$, onde $[a,b]\subset supp(X)$, isto é, parte das possíveis realizações de X simplesmente não é observada. Lidar com este caso também requer habilidade em modelar a truncagem. Em geral, se soubermos qual a classe a qual a distribuição $f(X;\theta)$ que deveria originar os dados pertence, a menos dos parâmetros θ (por exemplo, classe de distribuições normais, onde $\theta=(\mu,\Sigma)$ não é conhecido), e se soubermos também os limites de truncagem (a,b), podemos recuperar a distribuição original via

$$f_T(a<x<b)*[F(b)-F(a)]=f(x)$$

Se não soubermos nada sobre f, toda a inferência será baseada em f_T, e existe obviamente uma infinidade de possíveis distribuições f coerentes com f_T, com diferentes ponderações

dentro da parte do suporte fora do intervalo de trncagem (f genericamente não identificada).

Finalmente, temos as distribuições censuradas, que diferem das truncadas pelo fato de observarmos não apenas $f_T(a < x < b)$, mas também a proporção da amostra fora dos limites de truncagem. A diferença é sutil mas fundamental, pois neste caso podemos estimar conjuntamente $f_T(a < x < b), a$ e b, além de observarmos $F(b), F(a)$. Amostras censuradas surgem em dois contextos: (i) observo X se $a < X < b$ e observo o número de observações fora do intervalo (por exemplo, numa amostra aleatória de recém-contratados e seguida por 10 anos em observo a duração do emprego para aqueles que foram demitidos antes de 10 anos, e o número de pessoas para as quais a duração é maior que 10), ou (ii) observo $a < X < b$ e $F(a), F(b)$, se por exemplo eu começar a medir a vida útil de uma amostra de lâmpadas e fixar uma regra do tipo "parar de medir quando 80% das lâmpadas queimar".

3.6.1 Modelo Tobit

Suponha que uma determinada variável Y está relacionada aos observáveis, X e não-observáveis, U, via

$$y_i = x_i'\beta + u_i$$

e que seja válida a hipótese de que:

$$U|X \sim N\left(0, \sigma^2\right)$$

Numa base de dados sem censura, a estimação de (β, σ) resultaria de uma simples aplicação do método de MQO. O modelo Tobit surge quando em nossa amostra observamos

apenas a variável

$$y^* = \begin{array}{l} y, \text{ se y} > 0 \\ \\ 0 \text{ se y} < 0 \end{array}$$

Um caso típico é o da distribuição de salários, num modelo alternativo ao de seleção visto na última aula, onde da mesma forma observamos apenas os salários dos ocupados mas o que determina estar ou não empregado é a lucratividade das firmas, de modo que indivíduos muito pouco produtivos não encontram emprego simplesmente porque contratá-los não poderia ser vantajoso para as firmas. Assim, enquanto y captaria a produtividade marginal do trabalho, associada a características produtivas (X, U) dos indivíduos, y^*, a renda, só seria igual à produtividade do trabalho para os indivíduos minimamente produtivos.

Neste caso nossa função de regressão fica:

$$\begin{aligned}
E(y|x) &= E(y^*|x, y^* = 0)\Pr(y^* = 0) + E(y^*|x, y^* > 0)\Pr(y^* > 0) \\
&= E(y^*|x, y^* > 0)\Pr(y^* > 0) \\
&= E\left(x'\beta + \sigma\frac{u}{\sigma}|x, \frac{u}{\sigma} > -\frac{x'\beta}{\sigma}\right)\Pr\left(\frac{u}{\sigma} > -\frac{x'\beta}{\sigma}\right) \\
&= \Phi\left(\frac{x'\beta}{\sigma}\right)\left[x'\beta + \sigma E\left(z|z > -\frac{x'\beta}{\sigma}\right)\right] \\
&= \Phi\left(\frac{x'\beta}{\sigma}\right)\left[x'\beta + \sigma\frac{\phi\left(\frac{x'\beta}{\sigma}\right)}{\Phi\left(\frac{x'\beta}{\sigma}\right)}\right] \\
&= \Phi\left(\frac{x'\beta}{\sigma}\right)\left[x'\beta + \sigma\lambda\left(\frac{x'\beta}{\sigma}\right)\right] \\
&= \Phi\left(\frac{x'\beta}{\sigma}\right)x'\beta + \sigma\phi\left(\frac{x'\beta}{\sigma}\right)
\end{aligned}$$

Se estivermos interessados no impacto de um aumento de X_k sobre Y, temos:

$$\begin{aligned}\frac{\partial E(y|x)}{\partial x_k} &= \beta_k + \beta_k \frac{d\lambda\left(\frac{x'\beta}{\sigma}\right)}{d\left(\frac{x'\beta}{\sigma}\right)} \\ &= \beta_k \left[1 - \lambda\left(\frac{x'\beta}{\sigma}\right)\left(\frac{x'\beta}{\sigma} + \lambda\left(\frac{x'\beta}{\sigma}\right)\right)\right]\end{aligned}$$

Como o termo em colchetes é positivo, conclui-se que o sinal da derivada da função de regressão com respeito ao regressor k tem o mesmo sinal que β_k.

4 Parte II Análise de dados longitudinais

Notation 2 *$Z_i^t = (Z_{i1}, ..., Z_{it})$ representa a história de realizações de Z para o indivído i.*

Considere o modelo linear relacionando uma variável dependente y com uma ou mais explicativas, $(X, U) = (1, X_1, ..., X_K, U)$, onde X contém as variáveis explicativas observadas pelo econometrista, e U é uma variável (neste caso escalar) não-observável. No caso mais simples, supomos que a relação g é linear, isto é:

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + U$$

A estimação de um modelo deste tipo depende essencialmente das hipóteses que o analista está disposto a fazer a respeito do comportamento do componente não observável U, e de sua relação com os demais determinantes de y, X. A solução mais comum (mas não necessariamente correta) para o problema acima é supor $E(U|X) = 0$ (ou ainda mais forte: $U|X \sim N(0, \sigma^2)$), que permite a estimação consistente dos parâmetros β por mínimos quadrados, bem como da distribuição de U.

Infelizmente, a suposição $E(U|X) = 0$ é irrealista para grande parte das situações em que estaremos interessados. Dentre as possíveis complicações, com frequência nossos dados refletem uma situação de equilíbrio, onde as variáveis explicativas envolvidas não são completamente exógenas ao sistema, mas sim fruto de decisões dos agentes econômicos, que podem ou não levar em conta a decisão (posterior ou simultânea) envolvendo y. No exemplo canônico, y seria o (log) do salário de uma amostra de trabalhadores, e uma das explicativas seria a escolaridade. Se nossa base não permitir a mensuração de habilidades não-observáveis tais como inteligência, iniciativa, etc., então a variável U será de algum modo uma combinação de todas essas habilidades (por exemplo: $U = b_0 + b_1 h_1 + ... + b_L h_L$). Neste exemplo, é plausível imaginar que as mesmas variáveis que são diretamente precificadas na equação de salários afetem também a determinação do nível de escolaridade, se por exemplo pessoas mais inteligentes tiverem mais facilidade (e menor custo) de aprender, ou mesmo que isso não seja verdade, se escolaridade e inteligência forem insumos complementares na determinação da produtividade individual (ex: $U = b_0 + b_1 h_1 S$).

Como então lidar com situações onde $E(U|X) \neq 0$? Na primeira parte do curso, vimos uma possível solução, através da modelagem explícita do processo de determinação das variáveis explicativas observáveis que podem ser endogenamente determinadas, incluindo-se aí o papel de U na determinação desta variável.

Na segunda parte do curso, exploraremos outra possibilidade, válida para uma classe mais restrita de modelos (a linearidade será quase que necessária para a maioria das nossas conclusões), mas menos sujeita à especificação da forma como X é determinado como função

de U, onde a teoria pode não corresponder à realidade.

Inicialmente, suponha que haja uma decomposição de U em uma parte fixa no tempo, v, e umchoque temporal, ε:

$$U_{it} = v_i + \varepsilon_{it}$$

Para ser mais preciso, escrevemos o modelo como:

$$y_{it} = X'_{it}\beta + v_i + \varepsilon_{it}$$

No modelo acima, v_i é comumente chamado de efeito/componente não-observado, variável latente, heterogeneidade não-observável e, se o exercício envolver indivíduos, efeito individual ou heterogeneidade individual. Já o termo ε_{it} é denominado distúrbio ou erro idiossincrático.

4.1 Definições

Exogeneidade fraca Seja um modelo econométrico paramétrico, descrito pela distribuição conjunta

$$\begin{aligned} Q_t &= (Y_t, X_t) \\ Q_t &\sim f(Q_t|Q_{t-1};\theta) \end{aligned}$$

Em palavras, observamos um conjunto de variáveis Q_t para cada indivíduo num determinado ponto do tempo, e a relação de causalidade $Y_t = g_t(X_t, U_t;\theta)$ pode depender da

história pregressa de Q_t (em nosso caso Markoviano, apenas o período anterior importa) e de parâmetros θ. A distribuição f depende da distribuição de U.

Reescrevendo a distribuição f, temos

$$f\left(Q_t|Q_{t-1};\theta\right) = f\left(Y_t|X_t, Q_{t-1};\theta_y\right) f\left(X_t|Q_{t-1};\theta_x\right)$$

Exceto pela hipótese Markoviana de que do passado de Q apenas Q_{t-1} importa para explicar o comportamento de Q_t, nada mais foi assumido, e a igualdade acima é genericamente válida. $\theta_y, \theta_x \subseteq \theta$ são partições de θ.

Suponha que nosso interesse seja estimar φ_y. Se (i) $\varphi_y = g\left(\theta_y\right)$; e (ii) θ_y, θ_x puderem ser variados independentemente, isto é,

$$(\theta_x, \theta_y) \in \Theta_x \times \Theta_y : \theta_x \in \Theta_x, \theta_y \in \Theta_y$$

então dizemos que X_t é fracamente exógeno para θ_y. Em termos práticos, esta definição permite que estimemos φ_y usando informação apenas de $f\left(Y_t|X_t, Q_{t-1};\theta_y\right)$, já que neste caso $f\left(X_t|Q_{t-1};\theta_x\right)$ não será informativa sobre φ_y.

Exogeneidade estrita É em geral enunciada como

$$X_{it} \underline{\parallel} \varepsilon_{is}; \forall i, s$$

Em termos práticos, o principal resultado da exogeneidade forte é que:

$$\begin{aligned} E\left[y_{it}|X_{i1},...,X_{iT},v_i\right] &= E\left[y_{it}|X_{it},v_i\right] \\ &= X_{it}'\beta + v_i \end{aligned}$$

Esta hipótese diz que toda a correlação existente entre os determinantes observáveis e não-observáveis de y_t se encontra na correlação existente entre X_t e v. Uma forma mais rígida de exogeneidade estrita é $E\left[y_{it}|X_{i1},...,X_{iT}\right] = E\left[y_{it}|X_{it}\right] = X_{it}'\beta$, que claramente implica a hipótese anterior, sendo portanto mais forte e provavelmente menos realista. Em particular, $E\left[y_{it}|X_{i1},...,X_{iT}\right] = X_{it}'\beta + E\left[v_i|X_{i1},...,X_{iT}\right]$, e a menos que $E\left[v_i|X_{i1},...,X_{iT}\right] = E\left[v_i\right]$, as duas versões da exogeneidade estrita diferirão (basta haver correlação entre X e v para que difiram).

Num contexto linear, exogeneidade estrita implica que se:

$$\underbrace{B}_{(gxn)}Q_t + \sum_i \underbrace{C\left(i\right)}_{(gxn)}Q_{t-i} = U_t$$

$g \leq n$ relações comportamentais

então:

$$rank\left(B\right) = g$$

Causalidade de Granger Dizemos que X causa Y no sentido de Granger se defasagens de Y não predizem X:

$$f\left(X_t|Q_{t-1};\theta_x\right) = f\left(X_t|X_{t-1}, X_{t-2}, ..., X_0;\theta_x\right)$$

Em geral, essa propriedade é útil se estivermos interessados em previsão. Para estimação e inferência não ajuda muito.

Predeterminação Semelhante a exogeneidade forte, exceto que:

$$X_{it} \underline{\|} \varepsilon_{is}; \forall i, s > t$$

Agora, apenas os $\varepsilon's$ futuros são independentes de X_{it}.

Em termos de estimação, exogeneidade estrita e predeterminação tem forte implicação na formação de momentos amostrais que possam ser utilizados em aplicação do método dos momentos generalizados.

Painel balanceado versus desbalanceado Num painel balanceado, observamos Q_{it} para todos os indivíduos e períodos de tempo, ao passo que num painel desbalanceado, não. Dependendo da estrutura do problema, a presença de missing values para variáveis explicativas X pode não acarretar perda alguma de informação (no caso de X ser invariante).

Note que ao considerarmos o caso de painéis desbalanceados, estaremos sempre supondo que o motivo para o desbalanceamento (isto é, para o fato de que alguns indivíduos não são observados em todos os períodos da amostra) é não correlacionado com nenhuma das variáveis explicativas do modelo. O caso em que o desbalanceamento é correlacionado com

as variáveis explicativas leva a um problema de seletividade da amostra, e potenciais soluções devem incorporar explicitamente este aspecto do problema.

4.1.1 Exemplos de modelos em painel com efeitos não-observáveis

Avaliação de programas de requalificação profissional suponha que os salários individuais possam ser descritos por

$$\ln w_{it} = \alpha_t + x'_{it}\beta + \delta_1 (prog_{it}) + v_i + \varepsilon_{it}$$

onde (i, t) indexa indivíduos e tempo, respectivamente. O intercepto pode variar no tempo, e X são variáveis que afetam salários e potencialmente a participação no programa. Em princípio, nada impede que v_i seja também um dos determinantes da decisão de participar do programa (por exemplo, se capturar habilidade), levando a um problema de auto-seleção. Este primeiro dilema coloca a questão: será que poderíamos assumir que nossa variável explicativa, *prog* é não correlacionada com o erro e aplicar mínimos quadrados? Refletir sobre a hipótese mais apropriada a respeito do comportamento do não-observável é sempre a principal tarefa num xercício econométrico.

A segunda questão diz respeito à plausibilidade da hipótese de exogeneidade. Pode parecer razoável aceitar que realizações futuras de ε não afetem a decisão de participar do programa hoje, mas será que realizações passadas de ε também o são? Não seria razoável que justamente as pessoas que sofreram mais choques negativos de renda estão mais propensas a ingressar em um programa de requalificação?

Finalmente, o exemplo permite ainda investigar a terceira questão geralmente relevante neste tipo de exercício: será que o efeito de *prog* sobre salários se esgota em um período? E se

o programa durar mais de um período, será que o efeito se acumula ou aquilo que se adquire em um dado período tem magnitude própria? A resposta a estas perguntas deve indicar sobre a conveniência ou não de se incluir defasagens de *prog* na equação de salários, duração e/ ou permanência do indivíduo i no programa como variável explicativa, etc., caracterizando a dinâmica do efeito do programa sobre rendimentos.

Modelo de defasagens distribuídas de concessão de patentes (Hausman, Hall e Griliches(1984) O objetivo dos autores é relacionar investimento em pesquisa com obtenção de patentes:

$$patentes_{it} = \alpha_t + x'_{it}\beta + \delta_0 (RD_{it}) + \delta_1 (RD_{it-1}) + ... + \delta_5 (RD_{it-5}) + v_i + \varepsilon_{it}$$

Neste caso, os gastos tem um efeito cumulativo não-linear sobre o número de patentes conquistadas, e X inclui variáveis tais como o tamanho da firma (medido por vendas ou número de empregados), qualificação dos empregados, etc. v_i é neste caso característica não-observável da firma, que provavelmente está também associada à decisão de investimento em pesquisa, já que firmas com v elevado têm mais chances de ser bem-sucedidas e gerar retorno ao investimento. O interesse maior é sobre o padrão encontrado na sequência de $\delta's$, o que determina quando e como gastar num projeto (e mais importante: quando parar caso o projeto pareça inviável). Adicionalmente, as questões levantadas no exemplo anterior se mantêm. Por exemplo, as possibilidades de êxito de futuras patentes podem depender de patentes obtidas hoje (se estas forem insumo para a epsquisa futura); a lucratividade que

permitirá fazer futuros investimentos também pode depender do sucesso no presente, e assim por diante.

Variável dependente defasada Num modelo simples de dinâmica salarial, suponha que:

$$\ln w_{it} = \beta_1 \ln w_{it-1} + v_i + \varepsilon_{it}$$

O objetivo deste tipo de modelo é em geral determinar o grau de persistência salarial, uma vez controlado por efeitos não-observáveis (interpretados aqui como produtividade individual). Neste caso, exogeneidade estrita claramente não vale, e métodos que dependam desta hipótese não se aplicam. Este é tipicamente o caso de painéis dinâmicos, onde técnicas de efeitos fixos e aleatórios devem ser substituídas por estimações por método generalizado dos momentos calcadas em suposições de predeterminação.

4.1.2 Estimação por Mínimos Quadrados Empilhados

Suponha que no modelo

$$y_{it} = X'_{it}\beta + \underbrace{(v_i + \varepsilon_{it})}_{U_{it}}$$

a hipótese de que $E(X'_{it}U_{it}) = 0$ seja plausível. Uma vez que nossas hipótese básicas geralmente garantem que $E(X'_{it}\varepsilon_{it}) = 0$, a suposição realmente restritiva (e controversa) é $E(X'_{it}v_i) = 0$. De cara, essa hipótese não pode valer se por exemplo X incluir defasagens de y, já que por definição neste caso y e v estariam correlacionados em todos os períodos.

Ainda que $E(X'_{it}U_{it}) = 0$, os resíduos de nossa regressão estariam correlacionados (por

quê?), e uma matriz robusta deveria ser utilizada para a obtenção de variâncias dos estimadores de β (ainda que MQO seja consistente). Além disso, como a variância do erro não necessariamente decai com o espaçamento das observações, as condições necessárias à inferência baseada em T indo a infinito não valem, de modo que as propriedades assintóticas dos estimadores residem fortemente em N indo para infinito.

$$var\left(\widehat{\beta}\right) = \left(X'X\right)^{-1}\left(X'\widehat{u}\widehat{u}'X\right)\left(X'X\right)^{-1}$$

4.1.3 Modelo de efeitos aleatórios

Assim como nos mínimos quadrados empilhados, o modelo de efeitos aleatórios coloca o componente persistente não-observável no resíduo, supondo portanto que:

$$E\left(\varepsilon_{it}|X_i, v_i\right) = 0 \qquad \text{(RE. 1a)}$$

$$E\left(v_i|X_i\right) = E\left(v_i\right) = 0 \qquad \text{(RE. 1b)}$$

onde $X_i = (X_{i1}, ..., X_{iT})$. Evidentemente, se juntarmos as duas hipóteses temos $E\left(U_{it}|X_i\right) = 0$, mais forte portanto que $E\left(X_{it}'U_{it}\right) = 0$ requerida anteriormente. A primeira linha é simplesmente a hipótese de exogeneidade estrita de X com respeito a ε, já feita anteriormente. A segunda, que torna o modelo mais restrito, permite derivarmos algumas propriedades interessantes. Em particular, poderemos utilizar Mínimos Quadrados Generalizados para estimar consistentemente β, melhorando a eficiência do estimador.

Os dados empilhados agora podem ser descritos por $\mathbf{y}_i = \mathbf{X}_i'\beta + \mathbf{U}_i$, onde $\mathbf{U}_i = c_i\boldsymbol{\iota}_i + \boldsymbol{\varepsilon}_i$,

e $\boldsymbol{\iota}_i$ representa o vetor de 1's para observações correspondentes ao indivíduo i. A matriz incondicional de variância é $\Omega = E\left(\mathbf{U}_i\mathbf{U}_i'\right)$, e necessita da condição de posto:

$$posto\ E\underbrace{\left(X_i'\Omega^{-1}X_i\right)}_{KxK} = K \qquad \text{(RE. 2)}$$

Com as hipóteses acima,

$$\widehat{\beta} = \left(X_i'\Omega^{-1}X_i\right)^{-1}X_i'\Omega^{-1}y$$

Um caso particular segue sendo o mais utilizado em análises de efeitos aleatórios ainda hoje. As hipóteses, ainda mais restritivas, são as de que todos os termos não-observáveis são homoscedásticos, isto é, $E\left(v_i^2\right) = \sigma_v^2$ e $E\left(\varepsilon_{it}^2\right) = \sigma_\varepsilon^2$, além de $E\left(\varepsilon_{it}\varepsilon_{is}\right) = 0$, para todo s e t. Com isso, $E\left(U_{it}^2\right) = \sigma_v^2 + \sigma_\varepsilon^2$, e

$$\Omega_i = \begin{bmatrix} \sigma_v^2 + \sigma_\varepsilon^2 & \sigma_v^2 & \cdots & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + \sigma_\varepsilon^2 & & \vdots \\ \vdots & & \ddots & \sigma_v^2 \\ \sigma_v^2 & \cdots & \sigma_v^2 & \sigma_v^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

ou

$$\begin{aligned}
\Omega_i &= \sigma_v^2 \iota\iota' + \sigma_\varepsilon^2 I_T \\
&= \sigma_\varepsilon^2 \left(Q + \theta \left(I_T - Q\right)\right) \\
Q &= I_T - T^{-1}\iota\iota' \\
\theta &= \frac{\sigma_\varepsilon^2 + T\sigma_v^2}{\sigma_\varepsilon^2} = 1 + T\frac{\sigma_v^2}{\sigma_\varepsilon^2}
\end{aligned}$$

onde ι é um vetor de dimensão T composto por 1's, e I_T é a matriz identidade de dimensão T. A matriz de covariância pode finalmente ser escrita como:

$$\Omega = \begin{bmatrix} \Omega_i & 0 & \cdots & 0 \\ 0 & \Omega_i & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \Omega_i \end{bmatrix}$$

Ao invés de depender de $T(T-1)/2$ parâmetros, as hipóteses acima reduziram a 2 os parâmetros que caracterizam Ω. A correlação entre os resíduos de um mesmo indivído em dois pontos do tempo é constante neste caso: $\rho = \sigma_v^2/\left(\sigma_v^2 + \sigma_\varepsilon^2\right)$, e mede a importância relativa do componente persistente na formação do erro de um determinado período. Podemos formalizar as hipóteses necessárias para os resultados acima como:

$$E\left(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' | X_i, v_i\right) = \sigma_\varepsilon^2 I_T \qquad \text{(RE. 3a)}$$

$$E\left(v_i^2 | X_i\right) = \sigma_v^2 \qquad \text{(RE. 3b)}$$

A estimação do modelo se dá por Mínimos Quadrados Generalizados Factíveis. Suponha que tenhamos estimadores consistentes para $\sigma_v, \sigma_\varepsilon$, $\widehat{\sigma}_v, \widehat{\sigma}_\varepsilon$. Então, poderíamos construir $\widehat{\Omega} = \widehat{\sigma}_v^2 \iota\iota' + \widehat{\sigma}_\varepsilon^2 I_T$, e a partir disto:

$$\widehat{\beta} = \left(\sum_{i=1}^{N} X_i' \widehat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' \widehat{\Omega}^{-1} y_i \right)$$

A forma mais comum de obter $\widehat{\sigma}_v, \widehat{\sigma}_\varepsilon$ é em dois estágios. No primeiro estágio, estima-se o modelo por Mínimos Quadrados Empilhados (já que as hipóteses de efeitos aleatórios são suficientes para que as de MQE valham). A partir disto, temos uma estimativa de $\widehat{U}_{it}$ para cada indivíduo/ período, e podemos construir:

$$\widehat{\sigma}_u^2 = \widehat{\sigma}_v^2 + \widehat{\sigma}_\varepsilon^2 = \frac{1}{NT - K} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{U}_{it}^2$$

Se agora somarmos apenas os produtos cruzados dos resíduos, sabemos que estes não sofrem influência de ε, ou seja:

$$E\left(U_{it} U_{is}\right) = \sigma_v^2$$

e podemos assim construir

$$\widehat{\sigma}_v^2 = \frac{1}{NT\left(T-1\right)/2 - K} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \widehat{U}_{it} \widehat{U}_{is}$$

e finalmente fazer $\widehat{\sigma}_\varepsilon^2 = \widehat{\sigma}_u^2 - \widehat{\sigma}_v^2$.Na prática, ainda que raramente, pode ocorrer $\widehat{\sigma}_v^2 < 0$, o que deve ser interpretado como forte sinal de autocorrelação negativa forte de U_{it}, sugerindo que a hipótese RE. 3a não é válida. Em geral, esse problema desaparece com a inclusão de

dummies de tempo no modelo. Se ainda assim o problema persistir, outra forma de estimação deve ser tentada (por exemplo, Mínimos Quadrados Generalizados Factíveis Irrestritos).

Teste de hipótese usando variância robusta (método de White) Se não estivermos seguros de que RE.3 valha, ou se quisermos fazer inferência a partir de Mínimos Quadrados Empilhados, devemos usar uma matriz de variância que acomode a possibilidade de heterocedasticidade e autocorrelação, $E\left(U_i U_i'\right)$, que pode ser construída a partir de um estimador consistente de β via

(i) $\widehat{U}_i = y_i - \widehat{\beta}' X_i$,

(ii) $\widehat{S} = \frac{1}{NT}\sum_{it} \widehat{U}_{it}\widehat{U}_{it}'$,

(iii) $A = \frac{1}{NT}\sum_i X_i'\widehat{S}^{-1}X_i$,

(iv) $B = \frac{1}{NT}\sum_i X_i'\widehat{S}^{-1}\widehat{U}_i\widehat{U}_i'\widehat{S}^{-1}X_i$,

(v) $\widehat{V} = Avar(\widehat{\beta}) = \frac{A^{-1}BA^{-1}}{N}$.

Podemos a partir disto testar então a hipótese de que $R\widehat{\beta} = r$, usando a estatística de Wald:

$$\left(R\widehat{\beta} - r\right)' \left(R\widehat{V}R'\right)^{-1} \left(R\widehat{\beta} - r\right) \sim F_{p,q}$$

Mínimos Quadrados Generalizados Factíveis Após estimar um primeiro estágio de MQE, construímos

(i) $\widehat{\Omega} = N^{-1}\sum_{i=1}^{N} \widehat{U}_{i,MQE}\widehat{U}_{i,MQE}'$

(ii) $\widehat{\beta}_{MQGF} = \left(X'\widehat{\Omega}^{-1}X\right)^{-1}\left(X'\widehat{\Omega}^{-1}y\right)$

MQGF será consistente se RE.1 e RE.2 forem válidas e, se $E(UU'|X) = \Omega$ assintoticamente é tão eficiente quanto $\widehat{\beta}_{EA}$, mesmo quando RE.3 for válida. A razão pela qual podemos preferir usar o estimador de efeitos aleatórios ao invés do estimador MQGF é a significativa perda de graus de liberdade envolvida em MQGF (em particular, quando a dimensão N não é enormemente superior a T, pois Ω tem T(T-1)/2 elementos a serem estimados). Casos intermediários, por exemplo envolvendo a suposição de uma determinada estrutura autoregressiva para o termo não-observado, podem restringir Ω, tornando a perda de graus de liberdade menor, mas generalizar as hipóteses restritivas de RE.

Testando a presença de efeito não-observável persistente Num contexto de efeitos aleatórios, equivale a testar se $\sigma_v = 0$. Uma forma simples é testar se os resíduos são autocorrelacionados, já que sob as hipóteses do modelo apenas v_i causaria autocorrelação. Este teste tem pouco poder, fazendo com que frequentemente não se possa rejeitar a nula.

Usando diretamente o estimador de σ_v, construa:

$$z = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\widehat{U}_{it}\widehat{U}'_{is}}{\left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\widehat{U}_{it}\widehat{U}'_{is}\right)^2\right]^{1/2}}$$

e proceda com teste unicaudal de $\sigma_v > 0$, com $z \sim N(0,1)$.

4.1.4 Modelo de efeitos fixos

Suponha agora que a exogeneidade estrita de X com respeito a ε é válida, mas que não podemos afirmar que o mesmo valha para X com respeito a v. Neste caso, os métodos vistos até agora não poderiam ser utilizados, pois a correlação existente entre X e v tornaria nossas

estimações viesadas. Como resolver o problema neste caso?

Veremos nesta seção duas formas inspiradas no mesmo princípio para contornar a endogeneidade de X. Ambas utilizam implicações diretas do modelo teórico estimado, onde transformações da equação de regressão eliminam a dependência de v. Para isso, vamos lembrar nosso modelo e continuar assumindo que X é estritamente exógeno com respeito a ε:

$$\begin{aligned} y_{it} &= \beta X_{it} + v_i + \varepsilon_{it} \\ & \\ & \varepsilon_{it} \underline{\perp\!\!\!\perp} X_{is}; \forall i,t,s \\ & \\ & \varepsilon_{it} \text{ i.i.d.(i,t)} \end{aligned}$$

Estimação em primeiras diferenças Na primeira solução, chamada de modelo em diferenças, computamos as primeiras diferenças de todas as variáveis envolvidas. Vamos antes chamar de X_i^0 os elementos de X que não variam no tempo (sexo, raça, etc.), e de X_i^1 os que variam. O modelo fica:

$$\begin{aligned} y_{it} - y_{it-1} &= \beta^1 \left(X_{it}^1 - X_{it-1}^1 \right) + \varepsilon_{it} - \varepsilon_{it-1} \\ \widetilde{y}_{it} &= \beta^1 \widetilde{X}_{it}^1 + \widetilde{\varepsilon}_{it} \end{aligned}$$

Para as variáveis transformadas, $\left(\widetilde{y}_{it}, \widetilde{X}_{it}, \widetilde{\varepsilon}_{it} \right)$, vale a propriedade de exogeneidade estrita de $\widetilde{X}_{it}$ com respeito a $\widetilde{\varepsilon}_{it}$ (consequência da mesma exogeneidade assumida anteriormente), e do fato de ε_{it} ser i.i.d, conclui-se que $var\left(\widetilde{\varepsilon}\right) = 2\sigma_\varepsilon^2$. Evidentemente, podemos no modelo

transformado estimar β^1 e σ_ε^2 pelos métodos conhecidos de mínimos quadrados ordinários. Qual o ganho e qual a perda?

A principal vantagem deste método é não precisar usar a hipótese forte de que v_i e X_{it} são não correlacionados. Por esta estratégia, v_i simplesmente desaparece do modelo. Como vimos, na maioria das aplicações microeconométricas em cross-section a interpretação mais aceita para a presença de um componente aleatório no modelo é a de que representaria determinantes não-observados de y, tais como inteligência ou destreza numa equação de salário, habilidade gerencial numa equação de lucratividade das firmas, ou grau de aversão a risco e/ ou propensão a fugir do contrato, num modelo de concessão de crédito. Na estratégia acima, só precisamos assumir que este componente persistente não-observável é constante no tempo, deixando flutuações em torno desta constante para um choque aleatório (que na maioria das vezes é interpretado como um erro de expectativa ou simplesmente sorte). Problemas surgem quando o componente persistente também varia no tempo ou - bem mais complicado - quando esta variação depende das decisões dos agentes no presente (no caso, dos valores das variáveis endógenas do modelo no presente).

O principal custo para nós é que se nossa pergunta principal envolver a estimação de β^0, isto é, dos coeficientes associados aos regressores que *não* variam no tempo, então ao tomar as primeiras diferenças veremos que estes coeficientes desaparecem da equação impossibilitando sua estimação. Além disso, o intercepto nesta regressão deve ser interpretado como o coeficiente de uma tendência temporal linear em uma regressão em níveis, já que o intercepto em níveis também desaparece. Note contudo que a exigência de variação temporal não

precisa ser necessariamente satisfeita para toda a amostra, mas para ao menos parte dela (apenas as que apresentarem variação contribuirão para a estimação, de modo que se poucas pessoas preenchem este requisito o ceoficiente pode ser estimado mas com pouca acurácia). Um comentário importante é que podemos interagir atributos invariantes no tempo com dummies temporais e/ ou tendências temporais. Neste caso, ainda que não seja possível medir o coeficiente "estático" associado a esta variável, é possível saber se o efeito da mesma está crescendo ou não no tempo. Um exemplo comum é o diferencial salarial entre homens e mulheres, caso em que o painel não nos permite estimar um diferencial controlado em nenhum ponto do tempo, mas ainda assim podemos medir se este diferencial está crescendo ou não e a que taxa.

Adicionalmente, note que ao tomar a primeira diferença de nossas variáveis tivemos que descartar um período inteiro de observações, com consequente perda de graus de liberdade na estimação dos coeficientes.

Estimação do efeito em um indivíduo médio da população Nosso modelo de efeitos aleatórios pode ser estimado de dois modos distintos: No primeiro, podemos simplesmente usar MQGF após estimar em um primeiro estágio a matriz de covariâncias dos resíduos supondo a estrutura RE como sendo verdadeira. No segundo, construímos a função de verossimilhança a partir das hipóteses de que $\varepsilon|X, v \sim N(0, \sigma_\varepsilon^2)$ i.i.d., e de que $v|X \sim N(0, \sigma_v^2)$:

$$\begin{aligned}
\pounds_{it}\left(y, X|\beta, \sigma_\varepsilon, \sigma_v\right) &\propto f\left(y_{it}|X_{it}, \beta, \sigma_\varepsilon, \sigma_v\right) \\
&= N\left(X'_{it}\beta + \overline{v}, \sigma_\varepsilon^2 + \sigma_v^2\right) \qquad \text{(MLE1)} \\
&= f\left(y_{it}|X_{it}, v_i, \beta, \sigma_\varepsilon, \sigma_v\right) f\left(v_i|X_{it}, \beta, \sigma_\varepsilon, \sigma_v\right) \\
&= N\left[X'_{it}\beta + v_i, \sigma_\varepsilon^2\right] N\left(\overline{v}, \sigma_v^2\right) \qquad \text{(MLE2)} \\
\pounds\left(y, X|\beta, \sigma_\varepsilon, \sigma_v\right) &= \prod_{i=1}^{N}\prod_{t=1}^{T} \pounds_{it}\left(y, X|\beta, \sigma_\varepsilon, \sigma_v\right) \qquad (1)
\end{aligned}$$

A partir deste formato, podemos tanto estimar a totalidade dos parâmetros usando MLE 1, quanto estimar o efeito para um indivíduo médio da população, escolhendo os parâmetros que minimizam:

$$\prod_{i=1}^{N}\int_{supp(v)}\prod_{t=1}^{T} f\left(y|X, v, \beta, \sigma_\varepsilon, \sigma_v\right) f\left(v|X, \beta, \sigma_\varepsilon, \sigma_v\right) dv$$

Neste caso, contudo, σ_v não pode ser estimado, e deve ser normalizado (o padrão é 1).

Estimação por efeitos fixos Apesar de compartilhar exatamente as mesmas hipóteses que a estimação em primeiras diferenças, o método conhecido por efeitos fixos usa uma transformação diferente do modelo original. Inicialmente, note que da mesma forma que tomamos a primeira diferença do modelo para fazer nossa inferência no caso anterior, poderíamos ter tomado a diferença de todos com respeito ao primeiro período de observação dos agentes, ou ainda a diferença com respeito ao período intermediário, ou de um modo geral, com respeito a qualquer média ponderada das observações disponíveis de cada agente, que ainda

assim conseguiríamos desaparecer com o efeito fixo e estimar consistentemente os parâmetros relativos às características individuais que variam no tempo.

No caso dos efeitos fixos, escolhemos uma particular média para ser subtraída de cada linha de observação dos agentes: a média aritmética. Sabemos que cada agente é observado T vezes no painel. Se tomarmos a média de cada variável para cada agente, $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}; \overline{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$, teremos como resultado do modelo que:

$$\begin{aligned} \overline{y}_i &= \beta\overline{X}_i + v_i + \overline{\varepsilon}_i \\ y_{it} - \overline{y}_i &= \beta^1\left(X_{it}^1 - \overline{X}_i^1\right) + (\varepsilon_{it} - \overline{\varepsilon}_i) \end{aligned}$$

Evidentemente, para todas as variáveis fixas no tempo a média iguala o valor das mesmas em cada ponto do tempo, e portanto desaparecem quando tomamos a diferença (do mesmo modo que no casa anterior). A variância do resíduo é também $2\sigma_\varepsilon^2$, e ambos os casos correspondem a situações particulares do modelo geral

$$\overline{\overline{y}}_i = \beta\overline{\overline{X}}_i + v_i + \overline{\overline{\varepsilon}}_i$$

onde $\overline{\overline{z}}_i = \sum_{t=1}^{T} \pi_t z_{it}$, para uma variável genérica z ponderada por pesos $\pi_t : \sum_{t=1}^{T} \pi_t = 1$. No caso da estimação em diferenças, o peso integral é dado à observação imediatamente anterior, ao passo que na estimação por efeitos fixos $\pi_t = 1/T$. Em todos os exemplos, a

variância do resíduo é:

$$\begin{aligned} var\left(\varepsilon_{it} - \overline{\overline{\varepsilon}}_i\right) &= var\left[\varepsilon_{it} - \sum_{s=1}^{T} \pi_s \varepsilon_{is}\right] \\ &= var\left[\sum_{s\neq t}^{T} \pi_s \left(\varepsilon_{is} - \varepsilon_{it}\right)\right] \\ &= 2\sigma_\varepsilon^2 \sum_{s\neq t}^{T} \pi_s^2 \end{aligned}$$

e pode-se mostrar que atinge um mínimo se os pesos forem todos iguais ($\sum_{s\neq t}^{T} \pi_s^2 = T^{-1}$).

Da mesma forma que antes, ao tomar a diferença entre o modelo original e uma média ponderada das observações individuais perdemos sempre uma linha de observações, o que se traduz em perda de N graus de liberdade nas estimações. Se chamarmos $\widetilde{z}_{it} = z_{it} - \overline{z}_i$, teremos o modelo

$$\widetilde{y}_{it} = \beta^1 \widetilde{X}_{it}^1 + \widetilde{\varepsilon}_{it}$$

e novamente podemos estimar os parâmetros $\left(\beta^1, \sigma_\varepsilon\right)$ consistentemente por MQO.

Estimador intra-classes e entre-classes O estimador de efeitos fixos também recebe o nome de estimados intra-classes (within estimator) na literatura, em oposição ao estimador entre-classes (between estimator). A idéia é a de que no estimador intra-classes, a fonte de informação é a variação de características (dependentes e explicativas) dos indivíduos ao longo do tempo em torno de uma média individual (e portanto desprezando a variação existente entre as médias de diferentes indivíduos), ao passo que o estimador entre classes toma as médias dos indivíduos como unidade de análise, desprezando a variação observada

entre as observações de um mesmo indivíduo. Mantendo a notação anterior:

$$\begin{aligned}
\beta_W &= \beta_{FE} = \left(\sum_{i=1}^{N} \widetilde{X}_i' \widetilde{X}_i\right)^{-'} \sum_{i=1}^{N} \widetilde{X}_i' \widetilde{y}_i \\
\beta_B &= \left(\sum_{i=1}^{N} \overline{X_i}' \overline{X_i}\right)^{-1} \sum_{i=1}^{N} \overline{X_i}' \overline{y_i}
\end{aligned}$$

Sob as hipóteses de exogeneidade estrita de X com respeito a v e ε (e portanto com respeito a U), ambos os estimadores fornecem estimativas consistentes dos parâmetros desejados, mas se X for estritamente exógeno apenas com respeito a ε então o estimador entre-classes será viesado.

Usando um pouco de álgebra linear, sabemos que

$$\begin{aligned}
\overline{z}_i &= T^{-1} \sum_{t=1}^{T} z_{it} = T^{-1} \iota' z_i \\
\widetilde{z}_i &= z_i - \iota \overline{z}_i = \left[I - \frac{\iota\iota'}{T}\right] z_i
\end{aligned}$$

, o que significa que nossos estimadores podem ser reescritos como:

$$\begin{aligned}
\beta_W &= \left(\sum_{i=1}^{N} X_i' \left[I - \frac{\iota\iota'}{T}\right]' \left[I - \frac{\iota\iota'}{T}\right] X_i\right)^{-'} \sum_{i=1}^{N} X_i' \left[I - \frac{\iota\iota'}{T}\right]' \left[I - \frac{\iota\iota'}{T}\right] y_i \\
&= \left(\sum_{i=1}^{N} X_i' \left[I - \frac{\iota\iota'}{T}\right] X_i\right)^{-'} \sum_{i=1}^{N} X_i' \left[I - \frac{\iota\iota'}{T}\right] y_i \\
\beta_B &= \left(\sum_{i=1}^{N} X_i' \iota\iota' X_i\right)^{-1} \sum_{i=1}^{N} X_i' \iota\iota' y_i \\
&= \left(\sum_{i=1}^{N} X_i' \frac{\iota\iota'}{T} X_i\right)^{-1} \sum_{i=1}^{N} X_i' \frac{\iota\iota'}{T} y_i
\end{aligned}$$

onde na passagem da primeira para a segunda linhas usamos o fato de que $\left[I - \frac{\iota\iota'}{T}\right]$ é simétrica e idempotente.

Como última observação, o estimador de efeitos fixos produz estimativas idênticas às que seriam obtidas caso incluíssemos dummies individuais no modelo em níveis, ao invés de subtrair a média num modelo transformado (o que é um bom exercício para fazer em casa). A análise envolve construir uma matriz $Z = [X, D]$, onde D é um conjunto de N dummies individuais, e mostrar que $(Z'Z)^{-1} Z'y$ produz coeficientes associados a X^1 idênticos a β_W. Na demonstração, usamos o teorema de Frisch-Waugh-Lovell para regressões particionadas. Tal como antes, é possível ver que as variáveis constantes no tempo desaparecem, por serem perfeitamente colineares com as dummies individuais.

No entanto, suponha que tirássemos da regressão todos os atributos invariantes no tempo (inclusive o intercepto), e regredíssemos y nas variáveis restantes e dummies individuais. A pergunta é: os coeficientes das dummies têm neste caso alguma interpretação? A resposta é não. Em princípio, poderíamos imaginar que estes coeficientes estariam de algum modo capturando o próprio efeito fixo ou compinações deste com os parâmetros associados às características invariantes, mas este claramente não pode ser o caso uma vez que todas as propriedades dos estimadores obtidos estão suportadas pela análise assintótica fazendo N ir a infinito. Os parâmetros de efeitos fixos obviamente utilizam somente observações de um mesmo indivíduo, não satisfazendo portanto as condições assintóticas necessárias para a consistência do estimador. Dito de outro modo, suponha que v_i seja uma realização de $v \sim f(v)$, e que numa subamostra homogênea em características invariantes, usássemos o

percentil do coeficiente da dummy individual para inferir o valor de v. Esta não poderia ser uma estimativa consistente, pois se a amostra aumentasse e mantivéssemos o mesmo percentil, o valor estimado subiria ou cairia dependendo apenas do verdadeiro valor de v estar acima ou abaixo da mediana de f. A esta inconsistência dá-se o nome de problema dos parâmetros incidentais.

Variância assintótica Da análise assintótica convencional, segue que:

$$\begin{aligned}\sqrt{N}\left(\widehat{\beta}_{FE}-\beta\right) &= \left(N^{-1}\sum_{i=1}^{N}\widetilde{X}_i'\widetilde{X}_i\right)^{-'}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^{N}\widetilde{X}_i'\widetilde{\varepsilon}_i\right)\\ &= \left(N^{-1}\sum_{i=1}^{N}\widetilde{X}_i'\widetilde{X}_i\right)^{-'}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^{N}\widetilde{X}_i'\varepsilon_i\right)\end{aligned}$$

onde a segunda igualdade resulta da idempotência mostrada anteriormente. Consequentemente:

$$\begin{aligned}\sqrt{N}\left(\widehat{\beta}_{FE}-\beta\right) &\sim N\left(0,\sigma_\varepsilon^2\left[E\left(\widetilde{X}'\widetilde{X}\right)\right]^{-1}\right)\\ avar\left(\widehat{\beta}_{FE}\right) &= \frac{\sigma_\varepsilon^2}{N}\left[E\left(\widetilde{X}'\widetilde{X}\right)\right]^{-1}\end{aligned}$$

Para poder fazer inferência baseado em nosso estimador, a peça que falta é uma estimativa não-viesada de σ_ε, que pde ser obtida através de:

$$\widehat{\sigma}_\varepsilon^2=\frac{\sum_{i=1}^{N}\left(\widetilde{y}_i-\widehat{\beta}_{FE}'\widetilde{X}_i\right)'\left(\widetilde{y}_i-\widehat{\beta}_{FE}'\widetilde{X}_i\right)}{N\left(T-1\right)-K}$$

Aqui vale uma observação importante. Ao longo da análise de efeitos fixos, usamos duas hipóteses crucial, a exogeneidade estrita de X com respeito a ε e a independência de ε_{it} tanto entre diferentes indivíduos quanto entre períodos. Além destas, impusemos uma hipótese simplificadora que pode ser flexibilizada, de que $E(\varepsilon\varepsilon') = \sigma_{\varepsilon}^2 I_T$. Se tal suposição for falsa (mas as demais verdadeiras), ainda assim podemos estimar consistentemente os coeficientes de nossa regressão por efeitos fixos, mas temos que tomar cuidado com a variância. O mais comum é termos nestes casos autocorrelação dos resíduos, o que pode sugerir alguma persistência também no componente ε.

A forma de corrigir o problema é novamente utilizar uma matriz robusta:

$$avar\left(\widehat{\beta}_{FE}\right) = \left[E\left(\widetilde{X}'\widetilde{X}\right)\right]^{-1} E\left(\widetilde{X}'\varepsilon\varepsilon'\widetilde{X}\right)\left[E\left(\widetilde{X}'\widetilde{X}\right)\right]^{-1}$$

que pode ser estimada via

$$\left(\widetilde{X}'\widetilde{X}\right)^{-1}\left(\sum_{i=1}^{N}\widetilde{X}_i'\widehat{\varepsilon}_i\widehat{\varepsilon}_i'\widetilde{X}_i\right)\left(\widetilde{X}'\widetilde{X}\right)^{-1}$$

Assim como no caso de efeitos aleatórios, se a matriz de covariância do termo idiossincrático não for diagonal a variância robusta acima apenas corrige os desvios que devem ser usados conjuntamente com $\widehat{\beta}_{FE}$ para inferência, mas não suprime o fato de que $\widehat{\beta}_{FE}$ não é eficiente. O estimador eficiente será novamente uma versão de MQG do estimador de efeitos fixos:

$$\begin{aligned}
\widehat{\beta}_{FEGLS} &= \left(\sum_{i=1}^{N} \widetilde{X}_i' \widehat{\Omega}^{-1} \widetilde{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \widetilde{X}_i' \widehat{\Omega}^{-1} \widetilde{y}_i\right) \\
\widehat{\Omega} &= N^{-'} \sum_{i=1}^{N} \widehat{\widetilde{\varepsilon}}_i \widehat{\widetilde{\varepsilon}}_i'
\end{aligned}$$

onde $\widehat{\widetilde{\varepsilon}}_i$ é o resíduo obtido em um primeiro estágio usando $\widehat{\beta}_{FE}$ tradicional.

Finalmente, foi dito anteriormente que, sob a hipótese de que $E(\varepsilon\varepsilon') = \sigma_\varepsilon^2 I_T$, o estimador de efeitos fixos produz estimativas mais eficientes que o estimador de primeiras diferenças. Se esta hipótese não for verdadeira, este resultado não será necessariamente correto. Em particular, o estimador em primeiras diferenças produz melhores estimativas se os resíduos idiossincráticos apresentarem autocorrelação de primeira ordem. No caso de termos apenas 2 períodos, os dois estimadores são idênticos.

4.1.5 Comparando efeitos fixos e aleatórios

No modelo de efeitos aleatórios, vimos que a matriz de variância pode ser escrita como:

$$\begin{aligned}
\Omega &= \begin{bmatrix} \Omega_i & 0 & \cdots & 0 \\ 0 & \Omega_i & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \Omega_i \end{bmatrix} \\
\Omega_i &= \sigma_v^2 \iota\iota' + \sigma_\varepsilon^2 I_T
\end{aligned}$$

e que o estimador se resume a uma versão restrita de mínimos quadrados generalizados com

$$\widehat{\beta}_{RE} = \left(X_i'\Omega^{-1}X_i\right)^{-1} X_i'\Omega^{-1}y_i$$

É fácil mostrar que:

$$\Omega^{-1} = \begin{bmatrix} \Omega_i^{-1} & 0 & \cdots & 0 \\ 0 & \Omega_i^{-1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \Omega_i^{-1} \end{bmatrix}$$

e que:

$$\begin{aligned} \Omega_i^{-1} &= \varphi_1 \iota\iota' + \varphi_2 I_T \\ \varphi_1 &= -\frac{\rho}{(1-\rho)(1-\rho-T\rho)} \\ \varphi_2 &= \frac{1}{1-\rho} \end{aligned}$$

Voltando ao estimador de efeitos aleatórios e usando o fato de que $\frac{X_i'\iota}{T} = \overline{X}_i$ e de que $\overline{X}_i'\iota\iota'\overline{X}_i = \overline{X}_i'\iota\iota' X_i$, temos que:

$$\begin{aligned} X_i - \iota\overline{X}_i &= X_i - \iota\frac{\iota' X_i}{T} \\ \sum_{i=1}^{N} X_i'X_i &= \sum_{i=1}^{N} X_i'\left(I_T - \frac{\iota\iota'}{T}\right)X_i + \sum_{i=1}^{N} X_i'\left(\frac{\iota\iota'}{T}\right)X_i \\ &= \underbrace{\sum_{i=1}^{N}\left(X_i - \iota\overline{X}_i\right)'\left(X_i - \iota\overline{X}_i\right)}_{W_{XX}} + \underbrace{T\sum_{i=1}^{N}\left(\iota\overline{X}_i\right)'\left(\iota\overline{X}_i\right)}_{B_{XX}} \end{aligned}$$

$$\begin{aligned} X_i'\Omega^{-1}X_i &= \sum_{i=1}^{N} X_i'\Omega_i^{-1}X_i \\ &= \varphi_1 T \sum_{i=1}^{N} \frac{X_i'\iota\iota' X_i}{T} + \varphi_2 \sum_{i=1}^{N} X_i'X_i \\ &= \varphi_1 T B_{XX} + \varphi_2 \left(W_{XX} + B_{XX}\right) \\ &= \varphi_2 W_{XX} + \left(\varphi_1 T + \varphi_2\right) B_{XX} \end{aligned}$$

Similarmente, podemos rescrever $X_i'\Omega^{-1}y_i$ como

$$\begin{aligned} X_i'\Omega^{-1}y_i &= \varphi_2 W_{XY} + \left(\varphi_1 T + \varphi_2\right) B_{XY} \\ W_{XY} &= \sum_{i=1}^{N} \left(X_i - \iota\overline{X}_i\right)' \left(y_i - \iota\overline{y}_i\right) \\ B_{XY} &= \sum_{i=1}^{N} X_i' \left(\frac{\iota\iota'}{T}\right) y_i \end{aligned}$$

de modo que:

$$\begin{aligned} \widehat{\beta}_{RE} &= \left[W_{XX} + \theta B_{XX}\right]^{-1} \left[W_{XY} + \theta B_{XY}\right] \\ &= \left[W_{XX} + \theta B_{XX}\right]^{-1} \left[W_{XX}\widehat{\beta}_{FE} + \theta B_{XX}\widehat{\beta}_{BE}\right] \\ \theta &= \frac{\left(\varphi_1 T + \varphi_2\right)}{\varphi_2} \end{aligned}$$

Em linguagem matricial, o estimador de efeitos aleatórios é uma espécie de média entre o estimador de efeitos fixos (within, $\widehat{\beta}_{FE} = W_{XX}^{-1}W_{XY}$) e o estimador entre-classes (between, $\widehat{\beta}_{BE} = B_{XX}^{-1}B_{XY}$).

- Quando $\rho = 0$, temos $\varphi_1 = 0$ e $\theta = 1$, caso em que $\widehat{\beta}_{RE}$ é simplesmente nosso estimador de Mínimos Quadrados Empilhados.

- Quando $\rho = 1$, Ω_i se torna singular e Ω_i^{-1} não existe.

- Quando todos os regressores são fixos no tempo, $W_{XX} = 0$ e $\widehat{\beta}_{RE} = \widehat{\beta}_{BE}$

- Quando $T \to \infty$ e $\rho \neq 0$, temos que $\theta = 0$, e nesse caso $\widehat{\beta}_{RE} = \widehat{\beta}_{FE}$. Isso significa que testes de consistência do estimador de efeitos aleatórios tendem a rejeitar a inconsistência confirme o painel se torna mais longevo.

Finalmente, se estivermos interessados em testar a presença de covariância não nula entre o efeito fixo e as covariadas, podemos fazê-lo da seguinte forma:

Se a covariância for nula, então tanto o estimador de efeitos fixos quanto o estimador entre-classes devem fornecer estimativas não-viesadas do verdadeiro parâmetro β, mas se a covariância não for nula, apenas o estimador de efeitos fixos seria não-viesado:

$$\begin{aligned}
\widehat{\beta}_{FE} &= \beta + W_{XX}^{-1} \sum_{i=1}^{N} X_i' \left(I_T - \frac{\iota\iota'}{T} \right) U_i \\
\widehat{\beta}_{BE} &= \beta + B_{XX}^{-1} \sum_{i=1}^{N} X_i' \left(\frac{\iota\iota'}{T} \right) U_i
\end{aligned}$$

sob $H_0 : cov\left(\widehat{\beta}_{FE}, \widehat{\beta}_{BE}\right) = 0$. O teste se baseia então em estimar o modelo primeiramente usando efeitos fixos e depois o estimador entre-classes, e por fim testando a correlação entre os resíduos dos dois modelos.

4.2 Apêndice: o que faz o comando xtreg do stata

Considere o modelo linear de nosso interesse e suas variantes:

$$\begin{aligned}
y_{it} &= \alpha + X_i^{0\prime}\beta^0 + X_{it}^{1\prime}\beta^1 + v_i + \varepsilon_{it} \\
\overline{y}_i &= \alpha + X_i^{0\prime}\beta^0 + \overline{X}_i^{1\prime}\beta^1 + v_i + \overline{\varepsilon}_i \\
(y_{it} - \overline{y}_i) &= \widetilde{y}_{it} = \left(X_{it}^1 - \overline{X}_i^1\right)' \beta^1 + (\varepsilon_{it} - \overline{\varepsilon}_i) \\
(y_{it} - \theta\overline{y}_i) &= (1-\theta)\left(\alpha + X_i^{0\prime}\beta^0\right) + \left(X_{it}^1 - \theta\overline{X}_i^1\right)' \beta^1 + (1-\theta)\, v_i + (\varepsilon_{it} - \theta\overline{\varepsilon}_i)
\end{aligned}$$

Os estimadores vistos em classe são essencialmente regressões baseadas em uma das fórmulas acima. Por exemplo:

- Se acreditarmos que X é estritamente exógeno com respeito a v e ε, podemos estimar β simplesmente empilhando os dados e rodando mínimos quadrados (generalizados ou não, dependendo da crença na autocorrelação e heterocedasticidade de ε) na primeira equação

- Se além disso impusermos uma estrutura de correlações coerente com a hipótese RE.3, podemos estimar a quarta equação por mínimos quadrados (ordinários, após obter estimativas de θ a partir de MQE ou do estimador entre-classes em um primeiro estágio) e obter o estimador de efeitos aleatórios (alternativamente, poderíamos estimar $\widehat{\beta}_{RE}$ por máxima verossimilhança).

- Tomando as médias das variáveis dependentes e explicativas, e acreditando na exogeneidade estrita do X com respeito a v e ε, podemos rodar um estimador de MQO

na segunda equação e obter o estimador entre classes, $\widehat{\beta}_{RE}$

- Finalmente, se estimarmos a equação em desvios da média, tal como mostrado na terceira equação, MQO produz $\widehat{\beta}_{FE}$.

Duas questões surgem a partir das variações do modelo original expostas acima. A primeira questão é sobre a utilidade de usar o estimador entre-classes em qualquer situação, visto que sob as hipóteses necessárias para que este estimador seja não viesado, há outro estimador (MQE ou RE) mais eficiente disponível (exatamente por utilizar a variação temporal observada entre as variáveis envolvidas para cada indivíduo, adicionalmente à variação das médias entre indivíduos). A esse respeito, suponha que estimemos os modelos entre-classes e intra-classes e que estes difiram estatisticamente. Ora, sob as hipóteses de exogeneidade necessárias para estimar o entre-classes, ambos seriam consistentes, então uma possível explicação para a diferença encontrada seria a de que X é exógeno com respeito a ε, mas não a v, e que o estimador entre-classes não é consistente (e por isso diferente). Outra explicação é a de que o verdadeiro modelo seria especificado pela equação:

$$y_{it} = \alpha + \overline{X}_i'\gamma^0 + \left(X_{it}^1 - \overline{X}_i^1\right)'\gamma^1 + v_i + \varepsilon_{it}$$

Neste caso, o estimador entre-classes computa:

$$\overline{y}_{it} = \alpha + \overline{X}_i'\gamma^0 + v_i + \overline{\varepsilon}_{it}$$

enquanto o intra-classes computa:

$$y_{it} - \overline{y}_{it} = \left(X_{it}^1 - \overline{X}_i^1\right)' \gamma^1 + (\varepsilon_{it} - \overline{\varepsilon}_{it})$$

Neste caso, mesmo que a hipótese de exogeneidade estrita com respeito aos dois componentes não-observáveis seja válida, os estimadores intra-classes e entre-classes serão distintos. Um exemplo deste caso seriam compras de um determinado bem, y, onde X seria a renda. Bens supérfluos poderiam ser mais sensíveis a variações transitórias de renda, enquanto bens duráveis e/ ou necessários (ex: comida) seriam mais determinados pela renda permanente, melhor capturada pela média de renda ao longo do painel. Se nosso estudo tenta demonstrar este fato através da estimação de uma série de regressões para diferentes bens, em cada regressão não deveríamos esperar os coeficientes das duas rendas idênticos.

A segunda pergunta se refere à presença de variáveis que mudem sistematicamente ao longo do tempo mas não entre indivíduos (ex: idade, num estudo de coortes, ou variáveis macro tais como inflação, PIB, etc, num estudo de firmas ou famílias), e seu impacto nos estimadores entre-classes e intra-classes. Escreva o modelo:

$$\begin{aligned}
y_{it} &= \alpha + \delta Z_t + X_i^{0\prime}\beta^0 + X_{it}^{1\prime}\beta^1 + v_i + \varepsilon_{it} \\
\overline{y}_i &= \left(\alpha + \delta\overline{Z}\right) + X_i^{0\prime}\beta^0 + \overline{X}_i^{1\prime}\beta^1 + v_i + \overline{\varepsilon}_i \\
(y_{it} - \overline{y}_i) &= \delta\left(Z_t - Z_{t-1}\right) + \left(X_{it}^1 - \overline{X}_i^1\right)'\beta^1 + (\varepsilon_{it} - \overline{\varepsilon}_i)
\end{aligned}$$

Agora, constatamos que na estimação entre-classes é impossível distinguir $\delta\overline{Z}$ de α, pois ambas são capturadas pelo intercepto da regressão.

$\mathbf{R}^2$ Usamos as seguintes definições:

$$\begin{aligned}
\widehat{y}_{it} &= \widehat{\alpha} + X'_{it}\widehat{\beta} \text{ (be,re,MQE)} \\
&= \overline{y}_i + \left(X^1_{it} - \overline{X}^1_i\right)' \widehat{\beta}^1 + \widehat{\widetilde{\varepsilon}}_{it} \text{ (fe)} \\
\widehat{\overline{y}}_i &= \widehat{\alpha} + \overline{X}'_{it}\widehat{\beta} \text{ (be,re,MQE)} \\
&= \overline{y}_i - \left(X^1_{it} - \overline{X}^1_i\right)' \widehat{\beta}^1 - \widehat{\widetilde{\varepsilon}}_{it} \text{ (fe)} \\
\widehat{\widetilde{y}}_{it} &= \left(X^1_{it} - \overline{X}^1_i\right)' \widehat{\beta}^1 + \left(\widehat{\varepsilon}_{it} - \overline{\widehat{\varepsilon}}_i\right) \text{ (be,re,MQE)} \\
&= \left(X^1_{it} - \overline{X}^1_i\right)' \widehat{\beta}^1 + \widehat{\widetilde{\varepsilon}}_{it} \text{ (fe)}
\end{aligned}$$

$$\begin{aligned}
R^{2(overall)} &= corr^2\left(y_{it}, \widehat{y}_{it}\right) \\
R^{2(between)} &= corr^2\left(\overline{y}_i, \widehat{\overline{y}}_i\right) \\
R^{2(within)} &= corr^2\left(\widetilde{y}_{it}, \widehat{\widetilde{y}}_{it}\right)
\end{aligned}$$

O principal comentário aqui é o de que o R^2 natural de cada regressão varia: Nos estimadores e efeitos aleatórios e MQE, $R^{2(overall)}$ fornece o equivalente ao R^2 da regressão (o próprio, no caso de MQE). No caso de estimador entre-classes, $R^{2(between)}$ é o R^2 original da regressão, e para o estimador de efeitos fixos o R^2 da regressão é o within. Não surpreende, portanto, que em geral $R^{2(between)}$ seja maior no estimador entre-classes do que nos demais, ou que o $R^{2(within)}$ seja maior no estimador intra-classe, pois no primeiro caso o estimador é desenhado exatamente para maximizar $R^{2(between)}$, e no segundo, para maximizar $R^{2(within)}$.

Finalmente, vale notar que a transformação utilizada para recuperar y_{it} no modelo de efeitos fixos não é única. Em particular, sabemos que o modelo em níveis com dummies individuais produz exatamente os mesmos coeficientes que o modelo em desvios com respeito à média. Desse modo, uma forma alternativa de computar o R^2 seria rodar a regressão linear incluindo dummies individuais e medir a correlação entre o valor predito de y nesta regressão e o próprio y. Fazendo deste modo, é possível verificar que o $R^{2(overall)}$ cresce significativamente, e de certo modo podemos dizer que o $R^{2(within)}$ representa o ajuste do modelo de efeitos fixos desconsiderando a contribuição dos próprios efeitos fixos para a reta de regressão.

4.3 Painéis dinâmicos

Até o momento, estudamos o modelo:

$$\begin{aligned} y_{it} &= \alpha + \beta' X_{it} + U_{it} \\ U_{it} &= v_i + \varepsilon_{it} \end{aligned}$$

que pode ser identificado e estimado (i) por MQE/ MQGF/ RE se $X_{it} \underline{\parallel} U_{is}; \forall i, t, s$, ou (ii) por FE/ primeiras diferenças se $X_{it} \underline{\parallel} \varepsilon_{is}; \forall i, t, s$. Nesta seção, consideraremos o caso em que:

$$\begin{aligned} y_{it} &= \alpha + \gamma y_{it-1} + \beta_0' X_{it} + \beta_1' X_{it-1} + U_{it} \\ U_{it} &= v_i + \varepsilon_{it} \\ \varepsilon_{it} &= \rho \varepsilon_{it-1} + \epsilon_{it} \end{aligned}$$

Há neste caso duas novidades: a primeira é que agora os resíduos são serialmente correlacionados, e a segunda é que a variável dependente defasada surge entre as variáveis explicativas da regressão.

Inicialmente, note que

$$\begin{aligned} U_{it} &= v_i + \varepsilon_{it} \\ &= U_{it-1} + (\varepsilon_{it} - \varepsilon_{it-1}) \\ cov\left(U_{it}, U_{it-1}\right) &= \sigma_v^2 \end{aligned}$$

e que portanto a autocorrelação do resíduo por si só não prejudica a consistência dos estimadores de mínimos quadrados, se estiver garantido que X_{it} é estritamente exógeno com respeito a U_{is} para todos os "leads" e "lags". A presença de autocorrelação de ε da mesma forma não produz inconsistências na estimação de β, mas seus desvios precisam ser corrigidos para incorporar a autocorrelação (e se quisermos estimar β de forma eficiente, então devemos incorporar esta correlação e a respectiva matriz de covariâncias diretamente na estimação, via MQGF). Da mesma forma, correções de White (covariâncias robustas) e MQGF acomodam possível heterocedasticidade dos resíduos.

Com variáveis dependentes defasadas dentre as explicativas, é fácil ver que a exogeneidade estrita de X com respeito a ε não implica em exogeneidade estrita de y_{t-1} com respeito a ε. Em particular:

$$cov\left(y_{it-1}, \varepsilon_{it-1} | X_i^T, v_i\right) = \sigma_\varepsilon^2$$

e não vale portanto a exogeneidade estrita de (y_{it-1}, X_{it}) com respeito a ε, contrapartida da situação anterior com y_{t-1} entre os regressores.

A pergunta então é: a exogeneidade estrita de X tem neste caso algum poder de identificação e impõe alguma restrição nos coeficientes que seja útil para a estimação dos mesmos? A resposta é sim.

Inicialmente, note que se $X_{it} \underline{\perp\!\!\!\perp} \varepsilon_{is}; \forall i,t,s$, então

$$X_{it} \underline{\perp\!\!\!\perp} \Delta\varepsilon_{is}; \forall i,t,s$$

$$\Delta\varepsilon_{is} = \varepsilon_{is} - \varepsilon_{is-1}$$

e que portanto :

$$\begin{aligned} E\left(\Delta\varepsilon_{it}|X_i^T\right) &= 0 \\ &\Rightarrow \quad E\left(\Delta\varepsilon_{it}X_i^T\right) = 0 \end{aligned}$$

Reescrevendo o modelo, sabemos que:

$$\begin{aligned} \Delta\varepsilon_{it} &= \Delta y_{it} - \gamma\Delta y_{it-1} - \beta_0\Delta X_{it} - \beta_1\Delta X_{it-1} \\ 0 &= E\left(\Delta\varepsilon_{it}X_i^T\right) \Rightarrow \\ 0 &= E\left[\left(\Delta y_{it} - \gamma\Delta y_{it-1} - \beta_0\Delta X_{it} - \beta_1\Delta X_{it-1}\right)X_{is}\right]; \forall s \end{aligned}$$

Temos com isto KxT momentos que emergem como implicações diretas do modelo e que podem ser utilizados para a sua estimação pelo método generalizado dos momentos. Como

queremos estimar os K coeficientes relacionados aos $X's$ e mais o coeficiente da variável dependente defasada, e como perdemos um período de observações ao tomar a primeira diferença das variáveis, o modelo é exatamente identificado se nosso painel tiver ao menos 3 períodos, e é superidentificado se houver 4 ou mais períodos. Com 3 períodos, a estimação do modelo é um MQO (ou MQGF) da equação em primeiras diferenças.

Note contudo que há uma importante diferença em relação ao modelo anterior: com y_{-1} entre as variáveis explicativas, não mais podemos assumir que o conjunto das explicativas é estritamente exógeno com respeito a v e ε, de modo que toda e qualquer estimação deve envolver a eliminação (ou tratamento apropriado) para o efeito fixo na equação a ser estimada. Tomar primeiras diferenças é apenas uma das possíveis formas de fazê-lo.

Até agora, grande parte da nossa discussão envolveu hipóteses de exogeneidade estrita de X com respeito a um ou mais dos componentes não-observáveis. Nos casos de MQE e RE, X era efetivamente determinado fora do modelo, ao passo que no caso de FE e primeiras diferenças admitiu-se a possibilidade de que o elemento persistente do resíduo contribuísse na determinação do nível mas não da variação de X.

Num modelo mais ambicioso, contudo, a endogeneidade de X deveria ser considerada integralmente. Para ser mais preciso, assumiremos a partir de agora que X é pré-determinado (Wooldridge chama esta hipótese de exogeneidade sequencial) com relação a ε:

$$E\left(\varepsilon_{it}|X_i^t, y_i^{t-1}\right) = 0$$

o que significa que choques ε correntes e futuros não são correlacionados com valores passados

de y e valores presentes e passados de X, mas que choques passados podem ter influenciado escolhas a respeito do nível de X (e y_{-1}) atualmente observados.

Example 3 *Indivíduos aversos ao risco montam seus portfólios com base na utilidade esperada:*

$$\max_{\{a_{kt}\}_{k,t}} E_t \sum_{t=1}^{\infty} \delta^t \frac{c_t^{1-\gamma}}{1-\gamma}$$
$$s.t. c_t = y_{it} + \sum_{k=1}^{K} \widetilde{r}_{kt} a_{kt-1} - a_{kt}$$

A solução do problema pode ser expressa pelas equações de Euler:

$$E\left[\left(\frac{c_t}{c_{t-1}}\right)^{-\gamma} \widetilde{r}_{kt} | I_{t-1}\right] = 0$$

Instrumentos válidos neste caso são atributos $Z_{t-1} \in I_{t-1}$, pois estão em princípio correlacionados com X_t (neste caso c_t) mas não com o choque ε.

$$E\left[c_t^{-\gamma} \widetilde{r}_{kt} Z_{t-1}\right] = 0$$

A pergunta então é: como a hipótese de predeterminação se relaciona com a de exogeneidade estrita? Em termos de momentos:

$$\begin{aligned}
\Delta\varepsilon_{it} &= \Delta y_{it} - \gamma \Delta y_{it-1} - \beta_0 \Delta X_{it} - \beta_1 \Delta X_{it-1} \\
0 &= E\left(\Delta\varepsilon_{it} X_i^t\right) \Rightarrow \\
0 &= E\left[\left(\Delta y_{it} - \gamma \Delta y_{it-1} - \beta_0 \Delta X_{it} - \beta_1 \Delta X_{it-1}\right) X_{is}\right]; \forall s < t
\end{aligned}$$

ou seja, descartamos todos as relações de ortogonalidade anteriormente obtidas para $s \geq t$. Momentos adicionais podem ser obtidos se usarmos transformações de X (se X for predeterminado/ estritamente exógeno, e pos consequeência para algum para t,s valer $E\left(\Delta\varepsilon_{it} | X_{is}\right) =$

0, então não apenas $E(\Delta\varepsilon_{it}X_{is}) = 0$ mas também $E(\Delta\varepsilon_{it}g(X_{is})) = 0$, para qualquer função contínua g). Outra potencial fonte de momentos são hipóteses adicionais sobre a estrutura de correlação serial dos resíduos. Se por exemplo soubermos que $E[\Delta\varepsilon_{it}\Delta\varepsilon_{it-j}] = 0$ para algum j, então sabemos que:

$$E\left[\begin{array}{c}(\Delta y_{it} - \gamma\Delta y_{it-1} - \beta_0\Delta X_{it} - \beta_1\Delta X_{it-1}) * \\ (\Delta y_{it-j} - \gamma\Delta y_{it-j-1} - \beta_0\Delta X_{it-j} - \beta_1\Delta X_{it-j-1})\end{array}\right] = 0$$

Note que mesmo que o erro siga uma média móvel de ordem (finita) J, momentos adicionais serão obtidos para defasagens de ordem superior a J.

Estimação por Método Generalizado dos Momentos Como regra geral, lidamos com o problema de variável dependente defasada primeiramente tomando a primeira diferença do modelo (para encontrar uma forma reduzida que não contenha os efeitos fixos), e em seguida construindo momentos amostrais correspondentes aos momentos implicados pelo modelo teórico.

Nossas hipóteses de exogeneidade e predeterminação fornecem os momentos desejados, que podem ser somados a outros implicados por hipóteses de homocedasticidade ($E(v_i^2) = \sigma_v^2; E(\varepsilon_{it}^2) = \sigma_\varepsilon^2$), resíduos serialmente não-correlacionados ($E(\varepsilon_{it}\varepsilon_{it-j}) = 0$), entre outros.

Tanto nos momentos implicados pela predeterminação/ exogeneidade quanto pelos determinados pela ausência de autocorrelação, a forma geral encontrada é

$$E[(\Delta y_{it} - \gamma\Delta y_{it-1} - \beta_0\Delta X_{it} - \beta_1\Delta X_{it-1})Z_{it}] = 0$$

para alguma função Z dos dados, $W^T = \left(y^T, X^T\right)$. A partir deste ponto, a construção do estimador segue os mesmos passos de um estimador linear de variáveis instrumentais padrão, com Z sendo o instrumento e $\widehat{B} = \left[W'ZH^{-1}Z'W\right]^{-1} W'ZH^{-1}Z'y$, onde $W_{it} = (X_{it}, y_{it-1})$ e $B = (\beta, \gamma)$, para alguma matriz positiva-definida de ponderadores, H. Por esta razão, com frequência termos como "instrumento", "condições de ortogonalidade", entre outros estão presentes na literatura de estimadores lineares de painel por GMM.

Também dentro do espírito de variáveis instrumentais, a superidentificação do modelo pode ser acomodada pela escolha de uma matriz apropriada de ponderadores (que nada mais faz do que ponderar a importância que cada momento tem na obtenção do estimador. Uma vez que com mais equações do que incógnitas nada garante que $E\left(\Delta\varepsilon' Z\right) = 0$ para todo Z, esta matriz confere importâncias distintas para os desvios destes momentos da meta nula). A principal diferença surge quando se tenta definir eficiência neste contexto. Enquanto em grande parte dos problemas envolvendo variáveis instrumentais o pesquisador conhece (ou determina) qual o conjunto de instrumentos válidos, - e nesse caso é possível mostrar que há uma matriz de ponderação ótima que combina os instrumentos de forma a minimizar o erro quadrático médio do estimador - no caso dos modelos lineares em painel uma mesma condição de ortogonalidade em geral dá origem a uma vastidão de momentos (instrumentos) válidos. Por esta razão, quando se define "eficiência" no contexto de GMM o respectivo conceito é sempre *condicional em um conjunto de momentos determinados pelo investigador como sendo os mais relevantes para a estimação.* Para momentos lineares, a matriz de ponderação mais comum em estimadores de 1 estágio é $H = Z'Z = \frac{1}{N}\sum_{i=1}^{N} Z_i' Z_i$.

Ainda relacionado ao paralelismo existente entre estimadores de variáveis instrumentais e GMM, é possível construir testes análogos ao de Hausman para validade e superidentificação, assim como construir estimadores em dois estágios. No caso de 2 estágios, primeiramente utiliza-se uma matriz positiva-definida qualquer para obter estimativas consistentes de B, $\widehat{B}^*$, e a partir deste, constrói-se os resíduos:

$$\widehat{\Delta\varepsilon}_{it} = \Delta y_{it} - B'\Delta W_{it}$$

e a matriz de ponderadores,

$$\widehat{H} = \frac{1}{N}\sum_{i=1}^{N} Z_i'\widehat{\Delta\varepsilon}_{it}\widehat{\Delta\varepsilon}_{it}'Z_i$$

e por fim, utiliza-se esta matriz para estimar o novo valor de $\widehat{B}$. Um cuidado importante com a linguagem é que o estimador de 1 estágio é o análogo ao tradicional estimador IV, que pode ser derivado de duas formas distintas, uma delas sendo o 2sls (mínimo quadrado de 2 estágios)[6]. No caso de GMM, dois estágios não é o mesmo que 2-sls.

Finalmente, há um único momento que emerge diretamente da equação em nível, e que não requer instrumentos. Trata-se da simples esperança *incondicional* de U_{it} : $E\left[y_{it} - \gamma y_{it-1} - X_{it}'\beta\right] = 0$.

Heterogeneidade versus dependência intertemporal (state dependence) Em ambos os casos, os dados revelarão correlação entre os valores presente e passado da variável

[6] 1^o estágio: OLS de W em Z, coleta $\widehat{W}$.
2^o estágio: OLS de y em $\widehat{W}$.

dependente. Note que ao contrário da análise de séries temporais, a inferência neste caso é toda baseada com $N \to \infty$, sendo portanto dispensáveis as exigências de estacionariedade e ergodicidade comuns àquela literatura. Em particular, nenhuma restrição precisa ser imposta ao valor absoluto de γ neste caso. Enquanto dependência intertemporal resulta de algum mecanismo externo à relação entre y e seus determinantes contemporâneos que causa persistência intertemporal (e é passível de ser modeloado como um mecanismo gerador de persistência dentro do modelo), a autocorrelação gerada pela heterogeneidade é intrínsica ao indivíduo, e faz parte da determinação conetmporânea de y (e cuja detecção é sintoma de omissão de variáveis relevantes).

4.4 Pseudo-painel e cross-sections repetidas

Com frequência, os dados disponíveis estão organizados em uma sequência de cross-sections que não formam um painel. Se as hipóteses de exogeneidade estrita de X com respeito a v e ε forem válidas, podemos simplesmente rodar nossas regressões, de MQE ou efeitos aleatórios, nas cross-sections empilhadas, pois obteremos estimadores consistentes. Com respeito a um painel, a única diferença será na modelagem da autocorrelação dos resíduos, que terá que ser suposta nula (eventuais controles de heterocedasticidade, contudo, podem ser feitos).

Na presença de efeitos fixos, contudo, as estimativas se tornam inconsistentes, e estimadores de efeitos-fixos ou primeiras diferenças não são factíveis para corrigir o problema.

A solução proposta por Browning, Deaton e Irish (1985) e Deaton (1985) era a de modificar a unidade de análise de indivíduos para coortes de indivíduos. Uma coorte é definida como um grupo de indivíduos com pertinência fixa, e cuja elegibilidade para a coorte pode ser

identificada através das informações disponíveis. Um exemplo é a coorte dos nascidos entre 1965 e 1970. Para grandes amostras, pesquisas consecutivas produzirão amostras aleatórias de membros de cada coorte. Outras características inatas tais como sexo, raça, etc. podem ser utilizadas para ampliar a amostra e aumentar a acuidade da identificação de coortes.

Considere a regressão entre indivíduos:

$$\begin{aligned} y_{it} &= \beta' X_{it} + v_i + \varepsilon_{it} \\ E\left(\varepsilon_{it} | X_{it}\right) &= 0 \end{aligned}$$

Seja g uma variável aleatória que discrimina a coorte a que o indivíduo pertence, dadas suas características, Z_i. Tomando esperanças condicionais em g, o modelo acima fica:

$$E\left(y_{it} | g_i\right) = \beta' E\left(X_{it} | g_i\right) + E\left(v_i | g_i\right) + E\left(\varepsilon_{it} | g_i\right)$$

ou

$$y^*_{gt} = \beta' X^*_{gt} + v^*_g + \varepsilon^*_{gt}$$

A hipótese central é a de que a composição e o tamanho das coortes é estacionário, de modo que o efeito-fixo-coorte, v^*_g, seja constante no tempo (hipótese análoga à usada no caso anterior, de DID). Usando o recíproco amostral de médias populacionais, nosso modelo fica:

$$\overline{y}_{gt} = \beta' \overline{X}_{gt} + \overline{v}_g + \overline{\varepsilon}_{gt}$$

Qual o problema que surge nesse caso? Sabemos que uma parte (possivelmente significante) da variabilidade de y e X ocorre intra-coortes, e ao tomar médias, estamos ignorando essa variabilidade na estimação do efeito de X em y.

Mas pior que isso, a estratégia de coortes induz a um erro de mensuração que pode ser potencialmente grande, viesando as estimativas em direção a zero. O problema surge porque nossas hipóteses dizem respeito a momentos populacionais, $E(Z|g)$, ao passo que no exercício utilizamos contrapartidas amostrais, $\overline{Z}$. Se escrevermos:

$$\begin{aligned} \overline{y}_{gt} &= y^*_{gt} + \xi^y_{gt} \\ \overline{X}_{gt} &= X^*_{gt} + \xi^x_{gt} \end{aligned}$$

então os termos (ξ^y_{gt}, ξ^x_{gt}) capturam o erro de medida. A forma de resolver o problema é utilizar variáveis preditas para estimar a matriz de covariância dos erros de medida, e utilizar essa estimativa para corrigir o estimador dos coeficientes de interesse. Suponha inicialmente que as observações individuais satisfaçam o sistema:

$$\begin{aligned} y_{it} &= y^*_{gt} + e^y_{it} \\ X_{it} &= X^*_{gt} + e^x_{it} \end{aligned}$$

e que:

$$\begin{pmatrix} e_{it}^{y} \\ e_{it}^{x} \end{pmatrix} \sim iid \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma_0 & \sigma_{0x}' \\ \sigma_{0x} & \Sigma^x \end{pmatrix} \right]$$

Nesse caso, estimativas de e^y, e^x podem ser obtidas via

$$\begin{aligned} \widehat{e}_{it}^{y} &= y_{it} - \overline{y}_{gt} \\ \widehat{e}_{it}^{x} &= X_{it} - \overline{X}_{gt} \end{aligned}$$

Finalmente, podemos estimar conjuntamente os efeitos-fixos-coorte, v_g^*, e os coeficientes de interesse, β, via:

$$\begin{bmatrix} \widehat{v}_g^* \\ \widehat{\beta} \end{bmatrix} = \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \begin{bmatrix} d_g' d_g & d_g' \overline{X}_{gt} \\ \overline{X}_{gt}' d_g & \overline{X}_{gt}' \overline{X}_{gt} - \widehat{\Sigma} \end{bmatrix} \right)^{-1} \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \begin{bmatrix} d_g' \overline{y}_{gt} \\ \overline{X}_{gt}' d_g - \widehat{\sigma}_{0x} \end{bmatrix} \right)$$

Neste caso, as provas de consistência de $\widehat{\beta}$ consideram GT indo para infinito.

Collado (1997) propôs alternativamente tirar a primeira diferença das equações em coorte, para eliminar o efeito-fixo-coorte, e então lidar com o problema de erro de medida através do uso de variáveis instrumentais. Incluindo os erros de medida, nosso modelo fica:

$$\begin{aligned} \overline{y}_{gt} - \xi_{gt}^{y} &= v_g^* + \left(\overline{X}_{gt} - \xi_{gt}^{x} \right)' \beta + \varepsilon_{gt}^* \\ \Delta \overline{y}_{gt} &= \Delta \overline{X}_{gt}' \beta + \Delta w_{gt} \end{aligned}$$

com o problema de que, nesse caso, Δw_{gt} é correlacionado com $\Delta \overline{X}_{gt}$ e portanto precisa

ser instrumentalizado. Candidatos a instrumento são, em geral, defasagens suficientemente distantes do próprio $\overline{X}_{gt}$.

4.5 Modelo de diferenças em diferenças

O modelo de diferenças-em-diferenças tem como inspiração uma situação em que queremos analisar o impacto de uma política específica ocorrida em um determinado ponto do tempo, t*, e temos em mãos um painel com t = 1,...,T períodos tal que 1 < t* < T. Tipicamente, a política de interesse afeta apenas uma subamostra N* < N das observações disponíveis, e podemos portanto, construir uma variável dummy D que separa tratados e não-tratados pela referida política.

Formalmente, o objetivo é medir o impacto de um regressor específico:

$$D_{it} = \begin{cases} 1, \text{ se o indivíduo } i \text{ recebe o tratamento} \\ 0, \text{ caso contrário} \end{cases}$$

sobre uma determinada variável y, cuja determinação se dá da seguinte forma:

$$y_{it} = \phi D_{it} + \delta_t + v_i + \varepsilon_{it}$$

Modelo de efeitos fixos com tratamento binário Por simplicidade, iremos primeiro analisar o caso em que D é o único regressor da equação. No modelo acima, δ_t captura um efeito-fixo temporal e v_i mede o efeito-fixo individual. Assim como no caso dos painéis dinâmicos, o primeiro passo é tomar a primeira diferença:

$$\Delta y_{it} = \phi \Delta D_{it} + (\delta_t - \delta_{t-1}) + \Delta \varepsilon_{it}$$

O efeito de tratamento ϕ pode então ser consistentemente estimado via MQO (sem constante e com todas as dummies temporais).

Diferenças em diferenças Suponha agora que tenhamos apenas 2 períodos, e que o tratamento ocorra apenas no período 2 (o que significa que $D_{i1} = 0$ para todo i). Nosso modelo agora fica:

$$\Delta y_i = \phi D_i + \alpha + u_i$$

Agora temos duas opções para estimar ϕ: a primeira é rodar uma regressão de Δy em um intercepto e uma dummy indicando o tratamento.

A segunda forma envolve separar a amostra em duas: tratados e não tratados, e, para cada uma delas, estimar a média de y antes e depois de t* (nesse caso, nos períodos 1 e 2). Chamemos essas médias de $\overline{y}_t^{tr}$ e $\overline{y}_t^{nt}$, para tratados e não-tratados, respectivamente. Agora:

$$\widehat{\phi} = \Delta \overline{y}^{tr} - \Delta \overline{y}^{nt}$$

a técnica pode ser estendida ao caso de repetidas cross-sections, com a utilização de hipóteses mais fortes. Em particular, o uso de repetidas cross-sections deve nesse caso permitir identificar a elegibilidade para o tratamento, tanto antes de t*, quanto depois de t* para os não-tratados.

Hipóteses implícitas em estimadores DID A primeira hipótese em que devemos acreditar é a de que os efeitos temporais, δ_t, são comuns a todos os indivíduos. Se por exemplo y for renda e o perfil etário dos rendimentos diferir para homens e mulheres, então não devemos misturar em um mesmo exercício DID homens e mulheres.

A segunda hipótese é a de que as composições dos grupos de tratados e não-tratados é estável antes e depois do tratamento. No caso de dados em painel, a primeira diferença elimina o efeito-fixo (e a estabilidade da composição é tão mais crível quanto mais balanceado for o painel). No caso de uma sequência de cross-sections, ao tomar as médias teremos:

$$\begin{aligned}
\overline{y}_t^{tr} &= \phi + \delta_t + \overline{v}_t^{tr} + \overline{\varepsilon}_t^{tr} \\
\overline{y}_t^{nt} &= \phi + \delta_t + \overline{v}_t^{nt} + \overline{\varepsilon}_t^{nt} \\
\phi &= \left(\overline{y}_2^{tr} - \overline{y}_1^{tr}\right) - \left(\overline{y}_2^{nt} - \overline{y}_1^{nt}\right) + \\
&\quad \left(\overline{v}_2^{tr} - \overline{v}_1^{tr}\right) - \left(\overline{v}_2^{nt} - \overline{v}_1^{nt}\right) + u
\end{aligned}$$

e a consistência de ϕ dependerá de

$$plim\left[\left(\overline{v}_2^{tr} - \overline{v}_1^{tr}\right) - \left(\overline{v}_2^{nt} - \overline{v}_1^{nt}\right)\right] = 0$$

Covariadas A inclusão de covariadas neste caso pode ser feita de duas formas. A primeira consiste em dividir a amostra em células homogêneas nas covariadas, e rodar um DID em cada uma delas. Se por exemplo quiséssemos incluir controles para sexo, raça e região, nossas células seriam grupos de homens brancos no Nordeste, mulheres brancas no Nordeste,

homens negros no Nordeste,...,Mulheres Negras no Sul. A vantagem desta estratégia é que é plenamente flexível no impacto das covariadas e na interação destas com o tratamento que queremos investigar. A desvantagem é que é computacionalmente mais intensa, e, se parte das covariadas for contínua, técnicas não-paramétricas de interpolação serão necessárias para a implementação.

A segunda possibilidade parte da observação de que ϕ pode também ser estimado por MQO em uma regressão do tipo:

$$y_{it} = \beta_0 + \beta_1 D_{i2} + \alpha\delta_2 + \phi\delta_2 D_{i2} + \varepsilon_{it}$$

e, portanto, poderíamos incluir regressores fazendo:

$$y_{it} = (\gamma_0 + \gamma_1 D_{i2} + \alpha\delta_2 + \phi\delta_2 D_{i2})(1 + \beta' X_{it}) + \varepsilon_{it}$$

ou numa versão simplificada:

$$y_{it} = \gamma_0 + \gamma_1 D_{i2} + \alpha\delta_2 + \phi\delta_2 D_{i2} + \beta' X_{it} + \varepsilon_{it}$$

Exemplos de utilização de DID em trabalhos empíricos incluem:

- Impacto de mudanças de salário mínimo em New Jersey sobre o nível de emprego (Card e Krueger). Controle: Pennsylvannia (composição da força de trabalho similar).

- Efeito de imigração em massa de cubanos para Miami (crise dos balseiros) sobre empregabilidade e salários dos nativos (Card, 1990).Controle: LA, Tampa, Atlanta e Houston (grandes proporções de negros e hispânicos e tendências similares de emprego).

4.5.1 Logit em painel

Como veremos a seguir, o logit em painel é apenas uma aplicação do logit condicional analisado na primeira metade do curso.

Suponha que observemos indivíduos fazendo escolhas sobre um mesmo objeto ao longo do tempo, e que o problema decisório seja satisfatoriamente descrito por:

$$\begin{aligned} y_{it} &= \beta' X_{it} + v_i + \varepsilon_{it} \\ \varepsilon_{it} &\sim st.logistic, \text{ i.i.d.} \end{aligned}$$

e que ao invés de observarmos y_{it}, apenas informação sobre $y_{it}^* = 1\left(y_{it} > 0\right)$ esteja disponível.

A probabilidade de que em um dado ponto do tempo o indivíduo possua $y_{it}^* = 1$ é:

$$\Lambda\left(\beta' X_{it} + v_i\right) = \frac{\exp\left(\beta' X_{it} + v_i\right)}{1 + \exp\left(\beta' X_{it} + v_i\right)}$$

o que não permite aplicar diretamente máxima verossimilhança, pois não observamos v_i.

Seguindo a estratégia do logit condicional, construímos uma nova variável, $C_i = \sum_{t=1}^{T} d_{it}$, e condicionamos nossa função de verossimilhança em C_i. Obviamente, se $C_i = 0$ nós sabemos de imediato que em todos os períodos $d_{it} = 0$ com probabilidade 1, o que significa que condicionar em C explica plenamente o comportamento dessa subamostra que não pode desse modo contribuir para a estimação de β. Da mesma forma, se $C_i = T$ sabemos que todos os períodos foram de $d_{it} = 1$.

Tomando o exemplo com $t = 2$:

$$\begin{aligned}
&\Pr\left[d_{i1}=0 \;\&\; d_{i2}=1 | d_{i1}+d_{i2}=1, X\right] \\
= &\frac{\frac{1}{1+e^{\beta' X_{i1}+v_i}}\frac{e^{\beta' X_{i2}+v_i}}{1+e^{\beta' X_{i2}+v_i}}}{\frac{1}{1+e^{\beta' X_{i1}+v_i}}\frac{e^{\beta' X_{i2}+v_i}}{1+e^{\beta' X_{i2}+v_i}}+\frac{e^{\beta' X_{i1}+v_i}}{1+e^{\beta' X_{i1}+v_i}}\frac{1}{1+e^{\beta' X_{i2}+v_i}}} \\
= &\frac{e^{\beta' X_{i2}+v_i}}{e^{\beta' X_{i2}+v_i}+e^{\beta' X_{i1}+v_i}}=\frac{e^{\beta' X_{i2}}}{e^{\beta' X_{i2}}+e^{\beta' X_{i1}}}
\end{aligned}$$

e de modo análogo,

$$\begin{aligned}
&\Pr\left[d_{i1}=1 \;\&\; d_{i2}=0 | d_{i1}+d_{i2}=1, X\right] \\
= &\frac{e^{\beta' X_{i1}}}{e^{\beta' X_{i2}}+e^{\beta' X_{i1}}}
\end{aligned}$$

Obviamente, assim como o efeito fixo foi eliminado quando condicionamos nossas probabilidades em C, todos os regressores X que não apresentam variação temporal também são na prática eliminados. Para ver isso, separe $X_{it} = \left(X_i^f, X_{it}^v\right)$, o que significa que $\beta' X_{it} = \beta^{f\prime} X_i^f + \beta^{v\prime} X_{it}^v$. Nas expressões acima, $e^{\beta' X_{it}} = e^{\beta^{f\prime} X_i^f} e^{\beta^{v\prime} X_{it}^v}$, e, colocando $e^{\beta^{f\prime} X_i^f}$ em evidência, vemos que:

$$\begin{aligned}
\Pr\left[d_{i1}=0 \;\&\; d_{i2}=1 | d_{i1}+d_{i2}=1, X\right] &= \frac{e^{\beta^{v\prime} X_{i2}^v}}{e^{\beta^{v\prime} X_{i2}^v}+e^{\beta^{v\prime} X_{i1}^v}} \\
\Pr\left[d_{i1}=1 \;\&\; d_{i2}=0 | d_{i1}+d_{i2}=1, X\right] &= \frac{e^{\beta^{v\prime} X_{i1}^v}}{e^{\beta^{v\prime} X_{i2}^v}+e^{\beta^{v\prime} X_{i1}^v}}
\end{aligned}$$

Como fica então a função de verossimilhança condicional nesse caso? Se multiplicarmos ao longo de todos os indivíduos, i, sabemos que para aqueles com $d_{i1} = d_{i2} = 1$, teremos

$C_i = 2$ e a probabilidade condicional será 1 (assim como para aqueles com $d_{i1} = d_{i2} = 0$), e suas contribuições desaparecerão no produto. Ficaremos então com:

$$\pounds\left(\beta|C\right) = \prod_{\substack{i=1 \\ C_i \neq \{0,1\}}}^{N} \frac{\left(e^{\beta^{v\prime} X_{i2}^{v}}\right)^{z_i} \left(e^{\beta^{v\prime} X_{i1}^{v}}\right)^{1-z_i}}{e^{\beta^{v\prime} X_{i2}^{v}} + e^{\beta^{v\prime} X_{i1}^{v}}}$$

onde $z_i = 1\left(d_{i2} = 1|C_i = 1\right)$.